TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Division of Bioinformatics, Biotechnology Center TU Dresden

STRUCTURAL BIOINFORMATICS TO UNDERSTAND THE ORIGIN OF THE GENETIC CODE

STRUCTURAL MOTIF DETECTION IN AMINOACYL-TRNA SYNTHETASES

Florian Kaiser
Born on: 21st June 1991 in Oelsnitz/Vogtl.

DISSERTATION

to achieve the academic degree

DOCTOR RERUM NATURALIUM (DR. RER. NAT.)

Referee
Assoc. Prof. Dr. Peter R. Wills
Advisor
Prof. Dr. Dirk Labudde
Supervisor
Prof. Dr. Michael Schroeder

Submitted on: 26th June 2018
Defended on: 18th September 2018

Für Sara, für meine Familie.

ABSTRACT

One of the most profound open questions in biology is how the genetic code developed. The blueprints for proteins are encoded by triplets of nucleic acids, which in turn require proteins for interpretation and replication. The mere existence of this self-referencing system is a chicken-and-egg dilemma. Aminoacyl-tRNA synthetases are key players in the transfer of genetic information and reflect the earliest episode of life. These enzymes are responsible for loading tRNA molecules with the correct amino acid. Two protein superfamilies of aminoacyl-tRNA synthetases emerged, each responsible for ten amino acids. Despite sequence and structure similarity, the delicate balance between these superfamilies is manifested in two structural motifs, which were identified in the context of this thesis: the Backbone Brackets and the Arginine Tweezers. Both motifs realize constant ligand recognition and can be found in almost all protein structures of aminoacyl-tRNA synthetases.

In this thesis, I thoroughly characterized Backbone Brackets and Arginine Tweezers. The specific characteristics of these motifs require high-precision methods for their detection and analysis. However, existing algorithms do not feature an adequate computational representation of structural motifs at the atom level and the support of isofunctional residue mutations. In order to address these limitations, I designed the Fit3D algorithm for template-based and template-free detection of structural motifs. I show that proper computational representation of structural motifs is crucial and improves accuracy up to 26% for a benchmark dataset. Fit3D is a general-purpose tool for structural motif detection in high-resolution protein structure data. In conjunction with the accelerating progress in experimental methods, the demand for such tools will increase rapidly over the next years.

I applied Fit3D to structures of aminoacyl-tRNA synthetases to investigate whether Backbone Brackets and Arginine Tweezers are universal building blocks for ligand recognition, and to quantify structural changes upon ligand binding. While the Arginine Tweezers motif is exclusively found in aminoacyl-tRNA synthetases and paralogs, the Backbone Brackets seem to be a general pattern to recognize functional groups of certain ligands. The results show subtle differences in side chain orientation for one structural motif and a backbone shift for the other. This suggests a structural rearrangement to be a general mechanism in some aminoacyl-tRNA synthetases. The detailed level of these analyses would not have been possible without high-precision structural motif detection with Fit3D.

The results emphasize the importance of structural motifs, which consist of only a few residues, for the global function of the enzyme. Furthermore, the stunning conservation of the structural motifs located in the core domains of aminoacyl-tRNA synthetases suggests their presence in the earliest predecessors of these enzymes. Both motifs might have played a fundamental role in shaping the genetic code as we know it.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

aa:CP amino acid:[carrier protein] ligase

aaRS aminoacyl-tRNA synthetase

ABD anticodon binding domain

AlaRS alanyl-tRNA synthetase

AMP adenosine monophosphate

API application programming interface

ArgRS arginyl-tRNA synthetase

AsnRS asparaginyl-tRNA synthetase

AspRS aspartyl-tRNA synthetase

ATP adenosine triphosphate

BD2K Big Data to Knowledge

CP1 Connecting Peptide 1

CSA Catalytic Site Atlas

CSV comma-separated values

CysRS cysteinyl-tRNA synthetase

EC Enzyme Commission

EF-P translation elongation factor P

ES Enolase Superfamily

FN false negative

FP false positive

GlnRS glutaminyl-tRNA synthetase

GluRS glutamyl-tRNA synthetase

GlyRS glycyl-tRNA synthetase

HisRS histidyl-tRNA synthetase

ID insertion domain

IleRS isoleucyl-tRNA synthetase

JMH Java Microbenchmark Harness

LeuRS leucyl-tRNA synthetase

LysRS lysyl-tRNA synthetase

MetRS methionyl-tRNA synthetase

MMTF Macromolecular Transmission Format

MSA multiple sequence alignment

NOS Nitric Oxide Synthase

PDB Protein Data Bank

PheRS phenylalanine-tRNA synthetase

ProRS prolyl-tRNA synthetase

PSE position-specific exchange

PylRS pyrrolysyl-tRNA synthetase

RMSD root-mean-squared deviation

SerRS seryl-tRNA synthetase

SFLD Structure-Function Linkage Database

SVD singular-value decomposition

T1Cu type I copper binding site

ThrRS threonyl-tRNA synthetase

TN true negative

TP true positive

TrpRS tryptophanyl-tRNA synthetase

TyrRS tyrosyl-tRNA synthetase

UPGMA Unweighted Pair Group Method with Arithmetic Mean

ValRS valyl-tRNA synthetase

PUBLICATIONS

F. Kaiser[◕], S. Bittrich[◕], S. Salentin, C. Leberecht, V. J. Haupt, S. Krautwurst, M. Schroeder & D. Labudde, "Backbone Brackets and Arginine Tweezers delineate Class I and Class II Aminoacyl tRNA Synthetases"
PLOS Computational Biology, vol. 14, p. e1006101, Apr 2018
Contribution: FK co-designed the study, analyzed data, contributed to the writing process, and created the paper figures.
Thesis: Chapter 5 is based on this paper.
Note: Cover article of *PLOS Computational Biology* April 2018 issue.
doi:10.1371/journal.pcbi.1006101

F. Kaiser, A. Eisold & D. Labudde, "A Novel Algorithm for Enhanced Structural Motif Matching in Proteins"
Journal of Computational Biology, vol. 22, pp. 698–713, Jul 2015.
Contribution: FK conceived the study, designed and implemented the Fit3D algorithm, and contributed to the writing of the paper.
Thesis: Section 6.1 in Chapter 6 is based on this paper.
doi:10.1089/cmb.2014.0263

F. Kaiser[◕], A. Eisold[◕], S. Bittrich & D. Labudde, "Fit3D: a web application for highly accurate screening of spatial residue patterns in protein structure data"
Bioinformatics, vol. 32, pp. 792–794, Mar 2016
Contribution: FK conceived the study, implemented the back end of the web application, and wrote the paper.
Thesis: Section 6.1 in Chapter 6 is based on this paper.
doi:10.1093/bioinformatics/btv637

F. Kaiser & D. Labudde, "Unsupervised Discovery of Geometrically Common Structural Motifs and Long-Range Contacts in Protein 3D Structures"
IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. PP, Dec 2017
Contribution: FK conceived the study, designed and implemented the algorithm, and wrote the paper.
Thesis: Section 6.2 in Chapter 6 is based on this paper.
doi:10.1109/tcbb.2017.2786250

[◕] These authors contributed equally.

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisors Prof. Dr. Dirk Labudde and Prof. Dr. Michael Schroeder for their continuous support and guidance throughout the entire project. Special gratitude is owed to Prof. Dr. Dirk Labudde who gave me the possibility to participate in a stimulating work environment for this project.

Many thanks go to Alexander, Christoph, and Sebastian B. who provided continuous support and feedback over many years. They made my work enjoyable as both colleagues and friends. Together with the people from the Schroeder Group, Joachim and Sebastian S., they inspired me to bring this project to life. The calculations performed for this thesis would not have been possible without the kind support of Marcel.

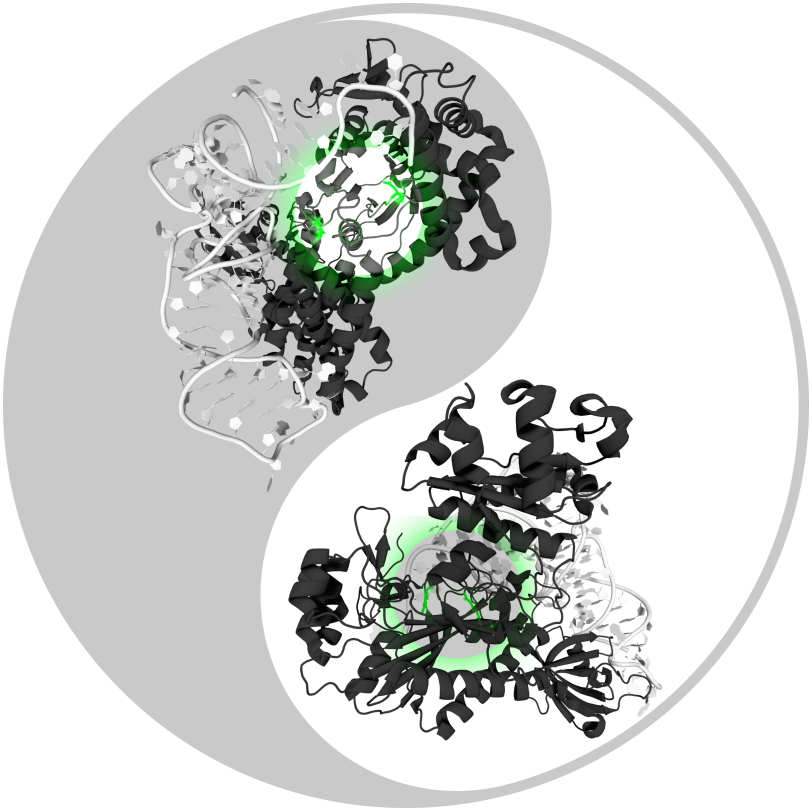
Moreover, I thank my family and my wonderful friends: my parents, Sara, David, and many others who supported me continuously during the last three years.

I am grateful for the funding of the project "HSMW Kooperative Promotionen 2015" by the European Social Fund (grant number: 100235472).

1. INTRODUCTION

1.1. MOTIVATION

The Code of Life Every known organism depends on the correct synthesis of proteins, which are composed of amino acid residues that are covalently linked and hence form a backbone-like construct. This linear construct folds into a complex three-dimensional structure *in vivo* which determines the molecular function of the protein [1]. Proteins realize a plethora of functions, of which many were studied in great detail, e.g. the degradation of peptides [2], but others remain enigmatic or completely unknown such as the function of the huntingtin protein [3]. The function and structure of a protein may be compromised by only minor changes in the correct succession of the respective amino acids [4–6]. This succession is encoded by nucleic acids. A complex molecular machinery of more than 100 proteins and nucleic acids is required to decode this information and to ensure efficiency and fidelity [7–10]. The ribosome is responsible to pair an mRNA codon with its corresponding anticodon of a tRNA molecule, which in turn delivers the cognate amino acid. The specification of the genetic code is hereby realized by aminoacyl-tRNA synthetases (aaRSs): a class of enzymes which ligate amino acids to their corresponding tRNA molecule [11]. Two different manifestations of aaRSs exist, each providing the building blocks of the other, which is why they constitute a unique self-referencing system.



Two manifestations of aaRSs exist that constitute a unique self-referencing system where each cannot exist without its counterpart. The idea for this image was inspired by Sebastian Bittrich and Alexander Eisold.

Genetic Code Emergence The mere existence of proteins and nucleic acids is a chicken-and-egg dilemma. The amino acid succession of each protein is encoded by nucleic acid blueprints. However, proteins are indispensable to replicate and translate nucleic acids. It is still unclear and heavily debated [12] how this reflexive system came to be [13] and which polymer type constituted the earliest primordial life forms.

The RNA world hypothesis assumes that nucleic acids were the sole basis of primordial life. RNA molecules can store and interpret genetic information, while also allowing for catalytic activity. In succession, proteins emerged to implement more elaborate, specific,

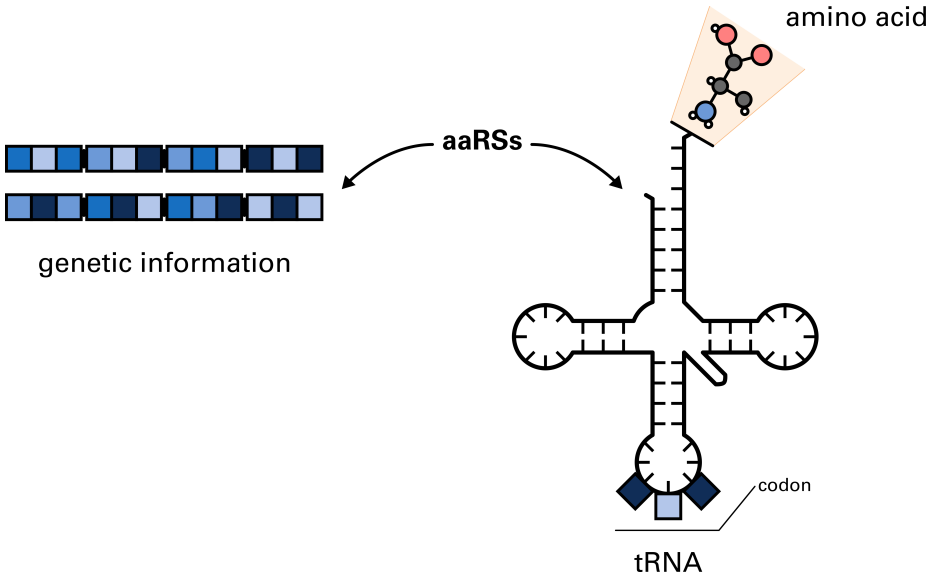


Figure 1.1.: The transfer of genetic information relies on the correct mapping between codons and amino acids. AaRSs are the enzymes which ligate amino acids to their corresponding tRNA and are thus the interface between gene and gene product.

and efficient catalytic activities [14]. However, the molecular complexity of RNA, its instability and limited catalytic repertoire [12, 15] raises concerns that such a primordial world was solely based on RNA.

Due to the less complex structure of simple peptides, and the chance of randomly generated catalytic polypeptides, there is the hypothesis that these catalytically active polypetides could have enriched without the intervention and coding by RNA [16, 17].

Another hypothesis assumes that genetic coding emerged from a system in which RNA and peptides coexisted and complemented each other from the very beginning [15, 18–21]. It is argued that only this interleaving of the two types of macromolecules can account for the speed with which the genetic code developed [19, 22–24]. Either way, aaRSs are the entities which most prominently reflect this early episode of life and are undoubtedly intertwined with the development of the genetic code.

Genetic Code Implementation In essence, the mapping of the genetic code is tied to the biochemical reaction catalyzed by aaRSs. AaRSs implement the mapping between codons and amino acids [15] (Figure 1.1) and are thus the key players in the transfer of genetic information. Beside the correct recognition of tRNA features [25], highly specific ligand interactions in the binding site are required to recognize the designated amino acid [11, 26–28] and to prevent errors in biosynthesis [26, 29]. These interactions are mediated by the correct arrangement of only a few residues within the protein. The detection of such key residues in the structure of a protein – so-called structural motifs – contributes to the understanding of protein structure [30], function [31, 32], and evolution [33]. Due to the large amount of known protein structures the manual detection of structural motifs is tedious or practically impossible. Consequently, computational methods are vital to understand how the catalytic mechanism in aaRSs, and hence the implementation of the genetic code, emerged.

Big Data in Bioinformatics Modern bioinformatics faces an increased demand for the processing of big data. With the rise of next-generation sequencing techniques and the advent of protein structure determination methods, terabytes of data are available for computational analysis [34]. The application of big data analytics has lead to breakthroughs in

biomedical research. Only the continuous development of innovative methods allows gaining new insights.

Deep learning algorithms were recently applied for the prediction of protein contact maps [35] or the assessment of protein model quality [36]. Furthermore, deep convolutional neural networks allow the prediction of the bioactivity of small molecules [37], which enables the design of new drugs in a structure-based manner. Huge public attention is drawn to these thriving fields of bioinformatics. The company Atomwise received \$45M[1] funding for their idea of exploiting established artificial intelligence algorithms for drug discovery.

With the Protein Data Bank (PDB) [38] as the largest data resource for macromolecular structures, a vast amount of data is publicly available for the analysis with computational methods. As of 2014, the PDB exceeded 100,000 structures. By now (March 2018[2]), the PDB contains over 134,000 entries and is growing at around 10% per year [39]. Most of the structure data is of atomic resolution and almost three quarters of the structures contain at least one ligand [1]. The high-troughput analysis of sequencing data allows the *in silico* prediction of even more protein structures [40]. The PDB includes approximately 1,000 structures of aaRSs across all kingdoms of life.

There is a high demand for general tools and algorithms to analyze this data that will increase rapidly in over the next years. The more protein structures are known, the more data will become available, which in turn allows gaining unprecedented insights. However, algorithms that exploit high-resolution protein structure data are scarce. Thus, this thesis aims at the development of general-purpose algorithms for the analysis of spatial data, which are then applied on protein structures of aaRSs to identify and characterize fundamental molecular mechanisms of these enzymes.

1.2. AIM AND OPEN PROBLEMS

Despite the availability of many aaRS structures in the PDB, these fundamentally important enzymes are mainly untouched in terms of structural motif analysis. With their implications for the origin of life, the overall aim of this thesis is:

> **Aim**
>
> A contribution to the understanding of the origin of the genetic code.

The high sequence and structure diversity of aaRSs poses a challenge for computational analyses. Due to the high degree of freedom of a protein's sequence, compared to its structure or function [4], sequence analysis can be error-prone. By using high-precision detection algorithms, subtle similarities as well as differences of structural motifs in aaRSs can be unveiled at the atom level. This is especially relevant for aaRSs because comprehensive structural analyses of these enzymes are scarce [22]. The characterization of elementary structural motifs in aaRSs can help to shed light on ligand recognition mechanisms, which were essential for the development of the genetic code during the earliest moments of life. Consequently, the first open problem which needs to be solved is:

> **Open Problem I**
>
> The identification and characterization of structural motifs in aaRSs (Figure 1.2).

[1] number taken from: atomwise.com/news, available as of March 20, 2018
[2] number taken from: rcsb.org/stats/growth/overall, available as of March 20, 2018

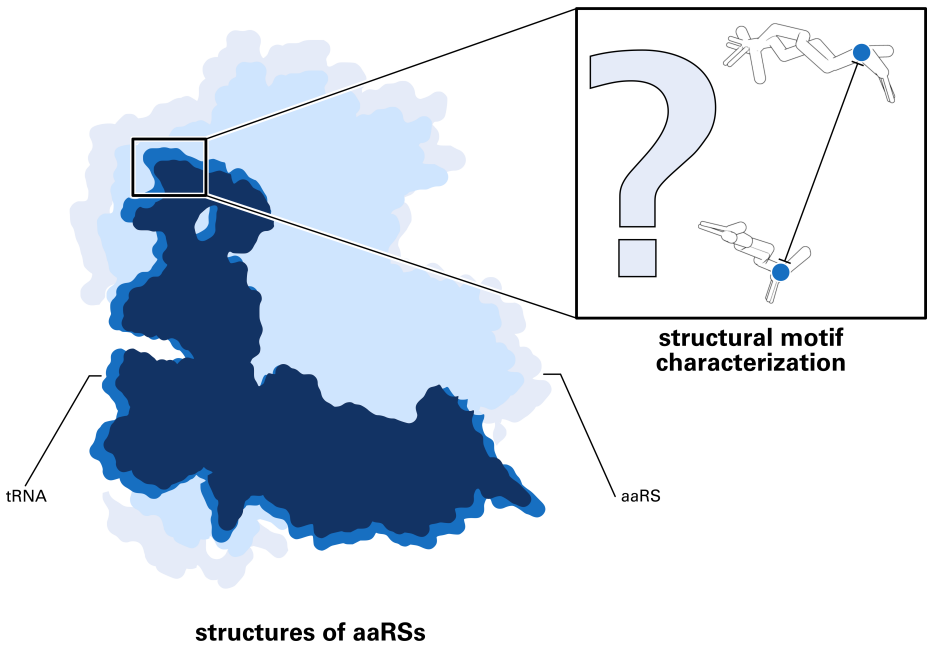


Figure 1.2.: The application of high-precision detection algorithms allows identifying and characterizing structural motifs in aaRSs. Based on the comparison of individual atoms or groups of atoms, functional implications can be deduced.

One of the major aspects for the computational analysis of aaRSs is the creation of a high-quality dataset of protein structures, derived from the PDB. Cross-references in biological databases, such as the Enzyme Commission (EC) number [41], can be used for the initial data acquisition. To characterize structural motifs in aaRSs, standardized alignment procedures are necessary which have to be implemented and tested for accuracy.

The consideration of full atomic resolution data is crucial in order to understand the molecular mechanisms in aaRSs at the greatest possible detail. Due to the high computational complexity of structural motif detection, existing algorithms are optimized for computational efficiency at the cost of accuracy. Nevertheless, structural motif detection algorithms are invaluable tools for the prediction of protein function [42, 43] or the detection of similar ligand binding sites for drug discovery [44]. The huge data source of protein structures allows to apply these algorithms at large scale. Thus, the second open problem which needs to be solved is:

Open Problem II

The development of general-purpose structural motif detection algorithms (Figure 1.3).

Due to the manifold biological roles of structural motifs, the problem of structural motif detection encompasses several biological aspects. Some structural motifs were shown to stabilize the global fold of the protein [30] and are not directly involved in protein function, whereas others are the functional determinants for ligand binding [45] or enzymatic reactions [2]. Due to these biological constraints, the required detection algorithms have to be as versatile and universal as possible and should not be limited to, for example, the detection of matches on the surface of a protein.

The computational aspects of structural motif detection require methods to handle the ever-increasing number of available protein structures. This can be achieved by using state-of-the-art data formats and the intelligent distribution of computational load (e.g. by parallel computation). Due to the complexity of the problem, which is related to subgraph isomor-

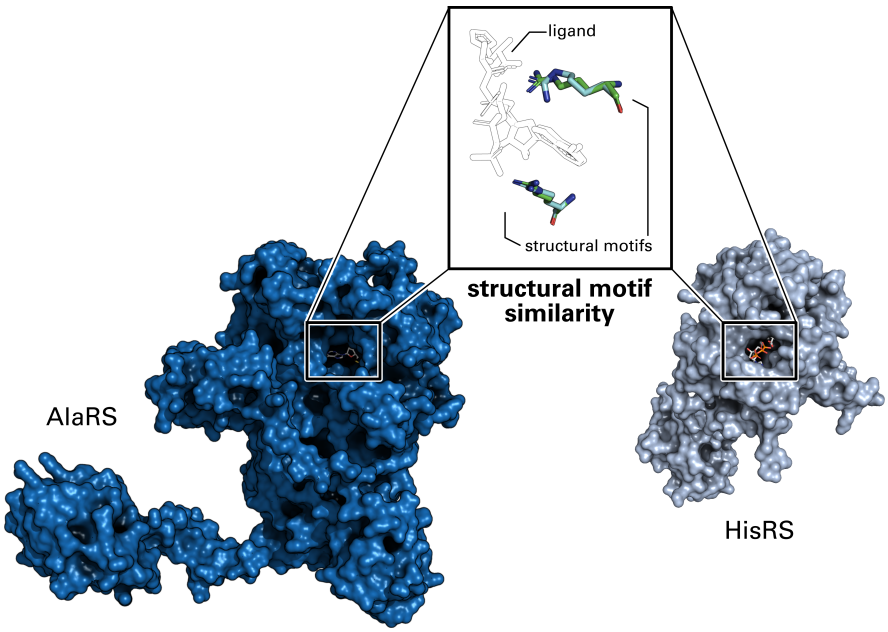


Figure 1.3.: The detection of local similarity in proteins at atomic resolution requires accurate matching algorithms. Even if the global structure similarity is low for a pair of proteins, they might share geometrically similar structural motifs. In the case of alanyl-tRNA synthetase (AlaRS) and histidyl-tRNA synthetase (HisRS) the overall sequence and structure similarity is low (≈15%). Nevertheless, the two proteins share a similar structural motif that is involved in ligand binding.

phism [46, 47] and pattern recognition, the application of heuristics and filtering steps is mandatory. Furthermore, the detection of similarity is usually based on a given template but a template-free approach is also required if input patterns are not available. Eventually, easy-to-use implementations of the developed algorithms have to be provided to the scientific community.

1.3. OUTLINE

Background Chapter 2 focuses on the biological role of aaRSs and how these enzymes relate to the origin of genetic coding and the protein biosynthesis in general. Sequence and structure studies of aaRSs are summarized. The hypothesis of ancient bidirectional coding of aaRSs enzymes is discussed alongside with the evolutionary separation of aaRSs into two distinct classes. The background Chapter 3 provides a foundation to understand the biological role of structural motifs in proteins. Various examples from literature are discussed, where structural motifs were shown to play a crucial biological role. The identification of structural motifs with computational methods is elucidated in Chapter 4. The problems of template-based and template-free structural motif detection are defined, state-of-the-art software and algorithms are reviewed. Based on this, limitations of existing methods are highlighted to substantiate the demand for new general-purpose structural motif detection algorithms.

Results The identified structural motifs in aaRSs are discussed in Chapter 5. This includes their characterization with the developed algorithms as well as their relation to the ancient forms of aaRSs. The chapter shows how structural bioinformatics algorithms can be applied to link evolution and genetic coding. The Fit3D algorithm is a comprehensive solution for the template-based and the template-free detection of structural motifs in macromolecular structure data. Chapter 6 explains the Fit3D algorithm in detail and highlights how Fit3D

addresses limitations of existing methods. Furthermore, the algorithm is validated and its algorithmic performance is analyzed. It is shown that the unique features of Fit3D increase the quality of structural motif detection. Chapter 7 summarizes the findings of this thesis and explains how the open problems were addressed.

Part I.

BACKGROUND

2. AMINOACYL-TRNA SYNTHETASES

This chapter incorporates content from the publication "Backbone Brackets and Arginine Tweezers delineate Class I and Class II Aminoacyl tRNA Synthetases" published in *PLOS Computational Biology*. For a detailed list of author contributions please refer to page 17.

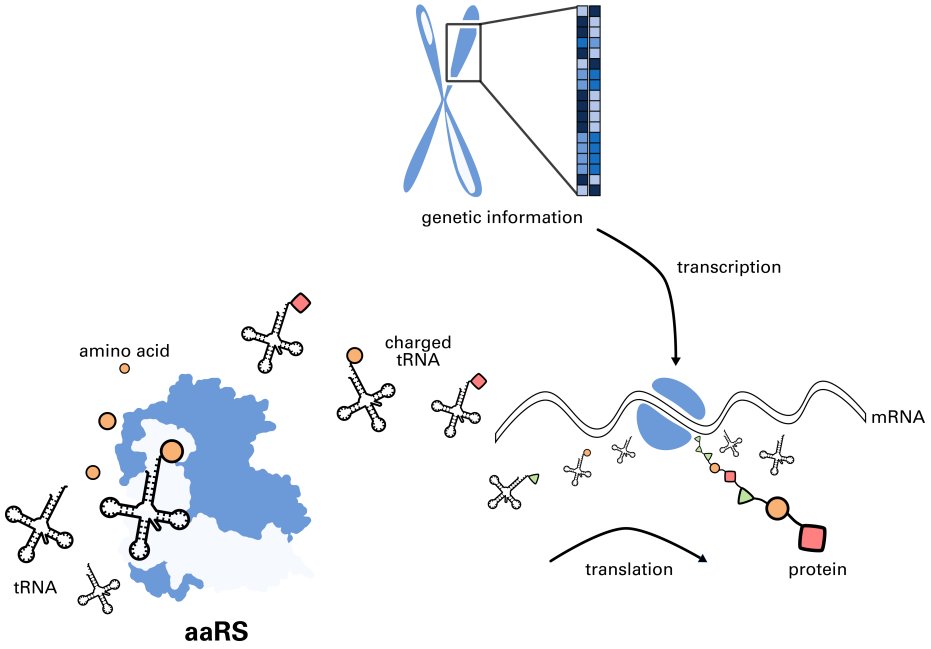


Figure 2.1.: The role of aaRSs protein biosynthesis. A specific aaRS is responsible to charge the tRNA with the correct amino acid. These charged tRNAs are then used by the ribosome to translate the final protein. The anticodons of the loaded tRNA molecules are matched against the mRNA codons which results in the correct succession of the amino acids of the protein.

2.1. BIOLOGICAL ROLE

Protein Biosynthesis AaRSs are fundamental enzymes for the correct translation of genetic information during protein biosynthesis (see Figure 2.1). After the transcription of genetic information by the RNA polymerase, the resulting mRNA is processed by the ribosome. The ribosome pairs an mRNA codon with its corresponding anticodon of a tRNA molecule that delivers the cognate amino acid. Once the amino acid cargo is removed, the tRNA molecules are released. Here, aaRSs are responsible to reload the free tRNA. These enzymes recognize the tRNA identity and catalyze the attachment of the amino acid to the corresponding tRNA molecule. Hence, the unique interface between gene and gene products is shaped by aaRSs [11, 22]. Three main theories have been proposed to explain the emergence of the self-encoding translational machinery, namely: coevolution [48], ambiguity reduction [49, 50], and stereochemical forces [51]. The interaction between amino acid and nucleic acid lies at the basis of each theory and is linked to the emergence of aaRSs [15, 52].

Enzymatic Reaction Every aaRS recognizes an amino acid and prevents misacylation of tRNAs by maximizing ligand specificity. The discrimination mechanisms between similar amino acids are well-studied [11, 26–28]. During the enzymatic reaction the designated amino acid is activated by consuming adenosine triphosphate (ATP), forming an aminoacyl adenylate, before it is linked to the cognate tRNA [53, 54]. For example, the fusion of aspartic acid and its corresponding $tRNA^{Asp}$ by the aspartyl-tRNA synthetase (AspRS) follows the two-step reaction:

$$\begin{aligned} \text{Asp} + \text{ATP} &\xrightarrow{\text{AspRS}} \text{Asp}{-}\text{AMP} + \text{PP}_\text{i} \\ \text{Asp}{-}\text{AMP} + \text{tRNA}^{\text{Asp}} &\xrightarrow{\text{AspRS}} \text{Asp}{-}\text{tRNA}^{\text{Asp}} + \text{AMP}. \end{aligned}$$

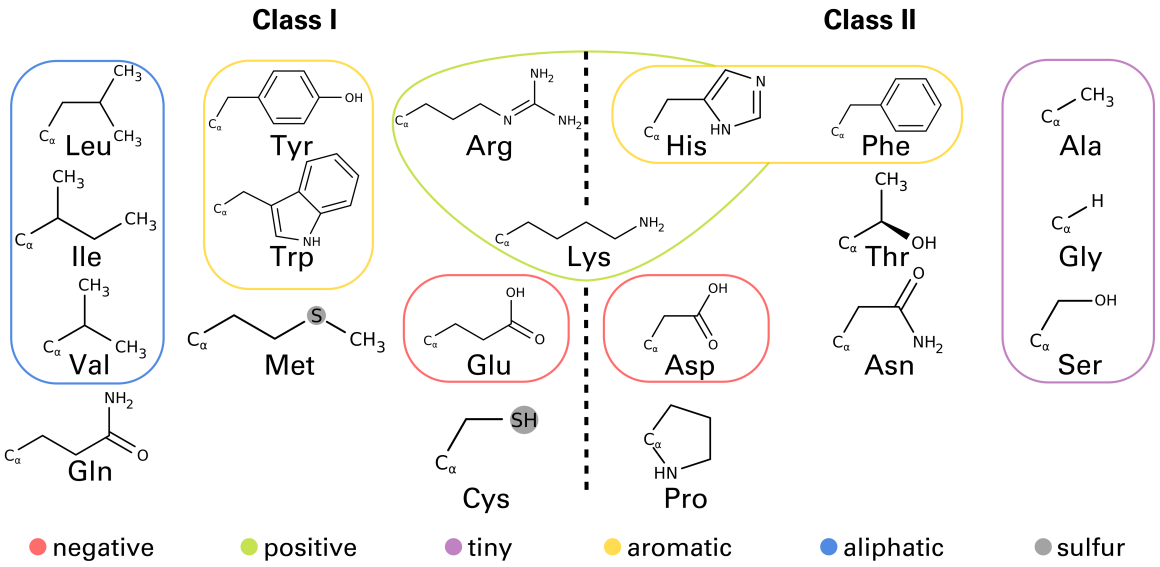


Figure 2.2.: Based on the physicochemical properties of the amino acids (colored according to [66]) no distinction can be made between the two aaRS classes. However, statistically significant differences based on amino acid side chain size [25] and binding site size [67, 68] are evident. Lysine is mostly processed by Class II aaRSs, but in all archaic organisms a Class I aaRS is responsible for lysine [69].

Today most organisms feature 20 concrete realizations, each handling one specific amino acid [55, 56].

Architecture The modular architecture of aaRSs has evolved well-orchestrated and was optimized for its specific requirements [57, 58]. Frequent domain inserts [19, 22] can render the evolutionary origin hard to track [59]. In principle, all aaRSs have to conserve three functions: the correct recognition of the tRNA identity and amino acid as well as the ligation of both. Commonly, the anticodon binding domain (ABD) ensures tRNA integrity by recognizing particular features of the anticodon [60, 61]. The identification and transfer of amino acids is then mediated by the catalytic domain, which differs in topology between the two classes. To minimize errors in protein biosynthesis, pre- and post-transfer editing mechanisms are conducted by approximately half of the aaRSs [26, 62, 63].

2.2. CLASSES AND TYPES

Sequence analyses revealed that aaRS enzymes can be divided into two complementary classes which differ significantly at the sequence and structure level. AaRSs have evolved divergently into Class I and Class II (Figure 2.2), where each is responsible for a distinct set of amino acids [58, 64, 65]. One concrete implementation of aaRSs is referred to as Type, e.g. arginyl-tRNA synthetases (ArgRSs) are the group of enzymes linking arginine to their corresponding tRNA. The physicochemical properties of amino acids are distributed evenly between both classes, even though amino acids handled by Class I were shown to be slightly bigger [25]. This suggest a concurrent emergence of both classes and that archaic aaRSs substrates have differed sufficiently to require two specialized kinds of aaRSs [22]. Both classes are, at several levels, as distinct as possible from each other [22] and share no sequence or structure similarity (Figure 2.3).

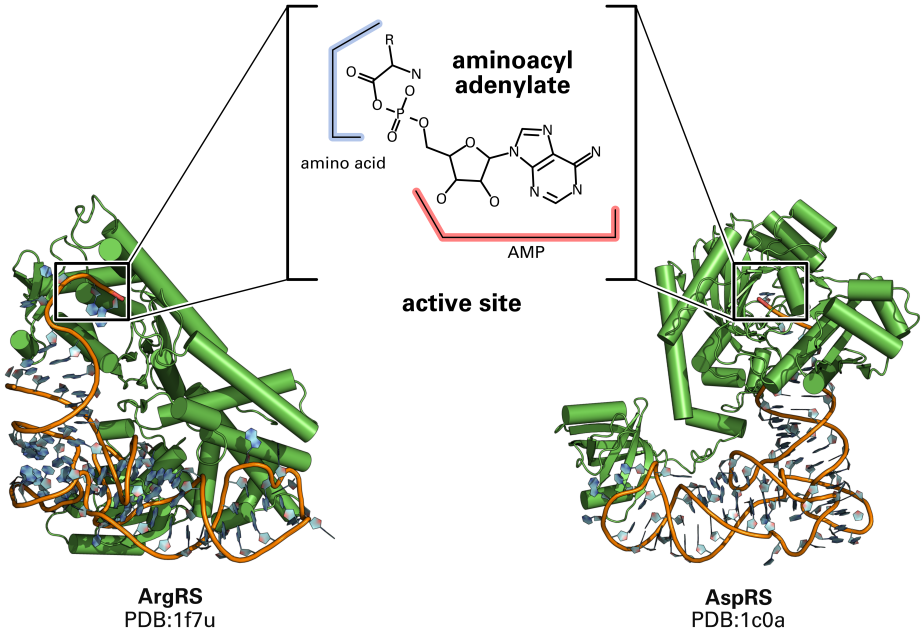


Figure 2.3.: Two structures of a Class I ArgRS (PDB:1f7u) and a Class II AspRS (PDB:1c0a), respectively. There is no similarity between the two classes at sequence or structure level. Prior to tRNA ligation, the amino acid ligand is converted to its activated form: aminoacyl adenylate, consisting of the amino acid and adenosine monophosphate (AMP).

2.3. SEQUENCE AND STRUCTURE

Sequence Motifs Sequences of aaRSs are highly diverse and result from fusion, duplication, recombination, and horizontal gene transfer [70, 71]. However, two sets of Class-specific and mutually exclusive sequence motifs have been identified, which are responsible for interactions with adenosine phosphate as well as catalysis [11, 64, 72]. Class I features the conserved HIGH and KMSKS motifs [11, 64]. The functional key motifs in Class II are referred to as Motif "1", Motif "2", and Motif "3" [11]. Both HIGH and KMSKS stabilize the transition state, whereby the latter constitutes a mobile loop in the folded structure [11]. The binding of ATP and the transition state of the reaction of individual Class I proteins have been demonstrated to be stabilized by a structural rearrangement [18, 73–80], which stores energy in a constrained conformation of the KMSKS motif [81]. The Class II motifs are less conserved [59] and more variable in their relative arrangement [64]. Motif "1" mediates the dimerization of protein structures, commonly found in Class II aaRSs [11, 82]. Motif "2" and "3" are essential for the reaction mechanism and feature two highly conserved arginine residues [64, 83, 84].

Structure Similarity The catalytic domain of Class I adapts the popular Rossmann fold [53], whereas Class II possesses a unique fold [71, 85, 86]. To assert the global structural similarity, two major structural alignments have been calculated for Class I and Class II, respectively, that revealed high structural similarity within each Class with average sequence identity below 10% [87]. At a functional level, both aaRS classes exhibit distinct ATP binding site architectures and reaction mechanisms. Class I aaRSs attach the amino acid to the 2'OH-group of the 3'-terminal adenosine of the tRNA, whereas Class II aaRSs use the 3'OH-group as the attachment location [88].

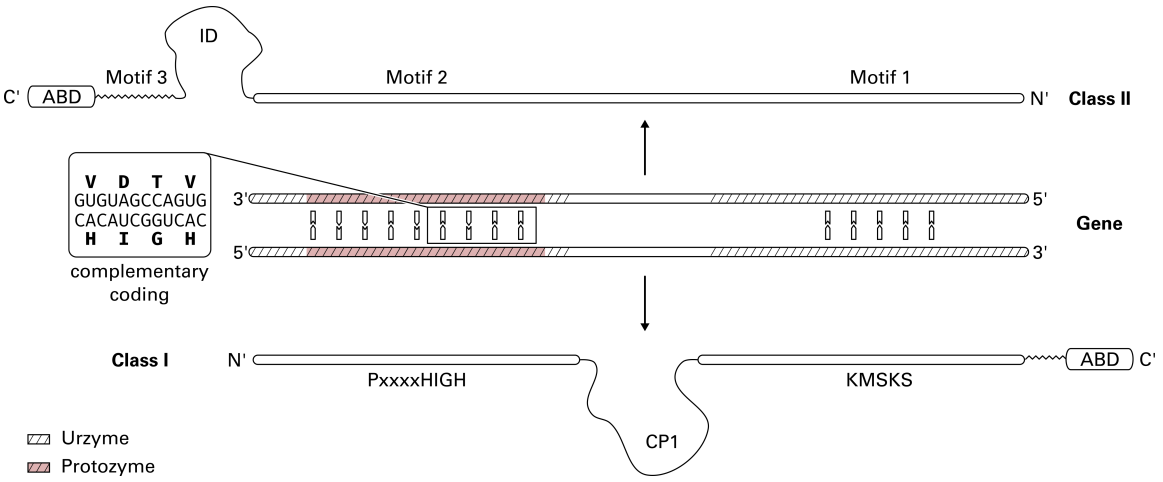


Figure 2.4.: The Rodin-Ohno hypothesis states that both aaRS classes descended from the opposite strands of a single gene. The signature motifs of each class were fully complementary on this gene. Both Protozymes originated from the complementary "HIGH-Motif 2" region (shaded in red). Contemporary aaRSs feature insertion domains (IDs), the Connecting Peptide 1 (CP1) as well as the addition of the ABD. Figure adapted from [19, 97].

2.4. COMPLEMENTARY BIDIRECTIONAL CODING

Rodin-Ohno Hypothesis There is strong evidence for two archaic proto-enzymes as the origin of all aaRSs, which were among the earliest proteins that enabled the development of life [18, 89–91]. In 1995, RODIN AND OHNO proposed an elegant explanation for the peculiarities that are observed in contemporary aaRSs: both classes were originally encoded on complementary strands of the same nucleotide fragment [18] (Figure 2.4). The Rodin-Ohno hypothesis is supported by an experimental deconstruction of aaRS sequences [19, 22]. In these studies, parts of contemporary aaRSs were removed and the catalytic strength of the resulting transcripts was assessed. One representative sequence of each Class was reduced to a peptide of only 46 amino acids. The coding nucleotide sequences of these 46-residue peptide were paired complementarily. These so called "Protozymes" were investigated regarding their structural and catalytic properties; they form molten globules [19, 22] and – despite the lack of ordered tertiary structure – they are still capable of rate enhancements by orders of magnitude [19, 22]. It is essential that the efficiency of different enzyme families across the proteome increases at comparable rates [19, 22]. The phenomenon of anti-parallel coupling of two genes was also postulated for other families of proteins [92, 93] and seems to be a phenomenon that affects the whole genome [94, 95]. One contradicting theory is the coevolutionary theory of the genetic code [48]. This theory suggests two main groups of amino acids based on the connectedness of their biochemical pathways and that amino acid biosynthesis was the dominant factor that shaped the genetic code [52]. Other authors suggested that both classes evolved from unrelated ancestors and are of independent origin [64].

The Rodin-Ohno hypothesis can explain why ATP and tRNA binding sites of both classes seem to be mirror images of each other [96] as well as the fact that both classes share virtually no similarities [11, 22, 71] beside their actual function [18, 19, 22]. All the contemporary aaRS Types are connected by the requirement to bind ATP. This basal unifying characteristic was found to involve hydrogen bonds in the Class I Protozymes [19].

Implications for the Genetic Code Remarkably, the restrictions inherent with a complementary coding may explain why the middle base of a codon is the most distinctive base for the corresponding amino acid nowadays [91]. Other studies showed how slight

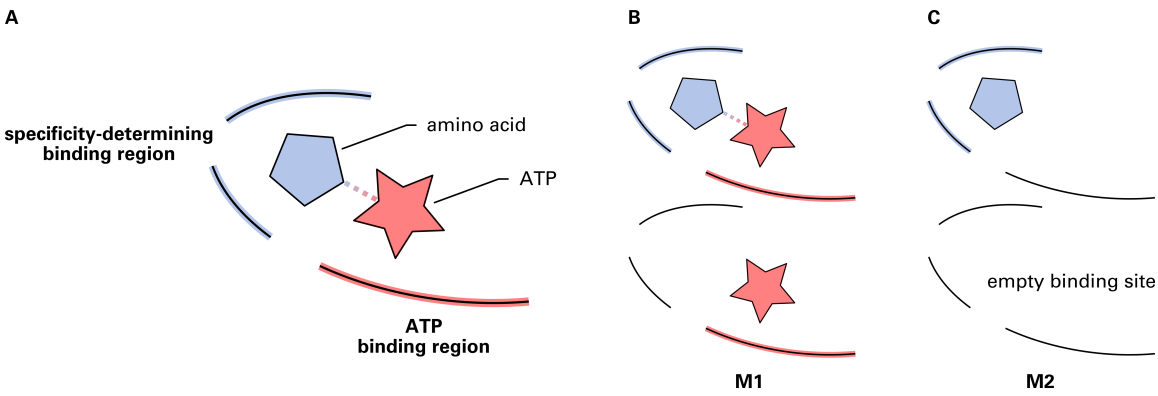


Figure 2.5.: The active site architecture of aaRSs and different binding modes based on the complexed ligand. (**A**) The active site of aaRSs can be divided into two parts. One part determines the specificity (blue shaded) of the enzyme by correct recognition of the designated amino acid ligand. The other part ensures the binding of ATP (red shaded). (**B**) Binding mode M1 is given if any ligands are present that bind to the ATP moiety of the binding site. This includes, for example, aminoacyl adenylate, ATP, or AMP. (**C**) In binding mode M2 either no ligands are present or only ligands that bind exclusively to the specificity-determining binding region.

differences in the substrate can result in a stable separation of aaRSs into two classes [15, 20]. Potentially, the two Protozymes diverged into ten aaRS Types each (Figure 2.2) and simultaneously increased fidelity and incorporated additional domains when necessary [18, 19, 89–91]. Most of aaRS evolution took place before the "Darwinian threshold" [87]. Only a few amino acids, such as tryptophan, were gradually incorporated into the genetic code after the last universal common ancestor and inefficient proteins evolved over time [52]. While similar amino acids were once processed by the same aaRS, specificity may have required additional aaRS Types to cope with increasing complexity. It is still possible to observe such generic aaRSs in some organisms [98, 99].

2.5. LIGAND BINDING CHARACTERISTICS

Despite their sequence and structure diversity, all aaRSs share one unifying aspect: the function of binding ATP as the necessary step for the activation of the enzyme's substrate. Hence, the most conserved part of the aaRS reaction mechanism is the amino acid activation with ATP, since it represents the principal kinetic barrier for the creation of peptides in a pre-biotic context [97]. This fundamental mechanism is shared by all Class I and Class II aaRSs, irrespective of their Type or the organism of origin.

The binding site of aaRSs can be divided into two major parts (Figure 2.5A): a specificity-determining region, where highly-specific ligand interaction realize the correct recognition of the designated amino acid [26], and the ATP binding region, which realizes constant binding of the ATP ligand within each Class [100]. Based on this division of the binding site, two binding modes can be defined: the state complexed with ATP (M1, Figure 2.5B) and the state in which no ATP is bound (M2, Figure 2.5C).

Furthermore, the catalytic domain has been predicted to constitute the ancestral precursors of aaRSs [19, 22, 93, 101] and is thus of outstanding interest to understand the origin of genetic coding.

3. STRUCTURAL MOTIFS

This chapter incorporates content from the publication "A novel algorithm for enhanced structural motif matching in proteins" published in *Journal of Computational Biology*. For a detailed list of author contributions please refer to page 17.

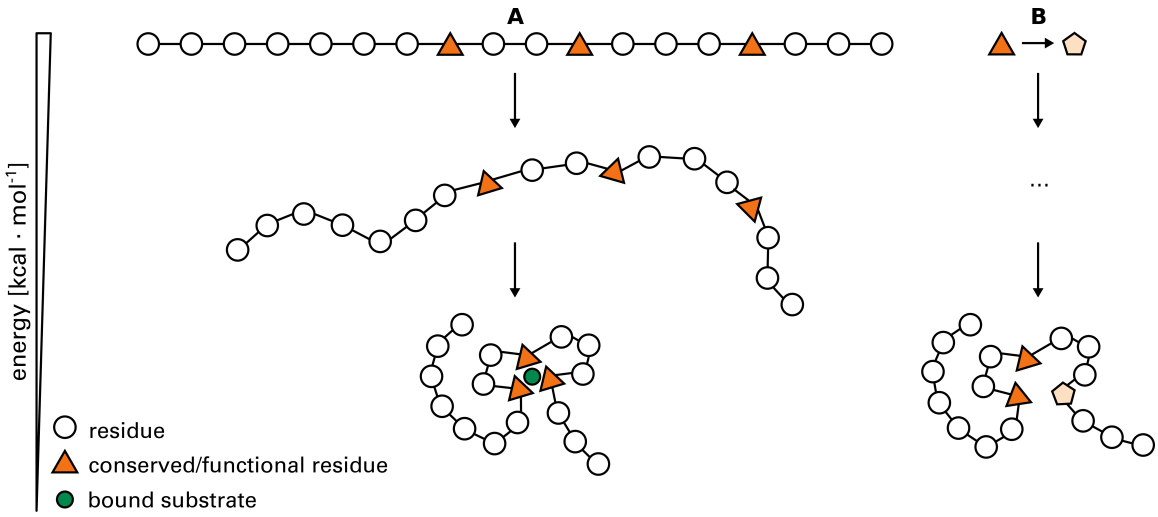


Figure 3.1.: The formation of functional "building blocks" during protein folding. (A) Highly conserved and functionally important residues are brought to spatial vicinity in the natively folded protein. The correct position of the residues relative to the substrate allows a biological function, e.g. the cleavage of peptide bonds. The residues depicted as orange triangles are considered to be a structural motif. (B) The mutation of a single residue leads to a disruption of the correct spatial assembly of the structural motif and subsequently to a loss of function such as the inability to bind or catalyze the substrate.

It is essential to identify and characterize structural motifs in the catalytic core domain of aaRSs to understand how ligand binding is realized. If universal structural motifs can be identified for both aaRS classes, this can give clues on how the formation and diversification of the genetic code took place in ancient times. The following chapter forms the basis to understand how structural motifs relate to protein function and evolution, and how these spatial residue patterns can be detected in datasets of protein structures.

3.1. PROTEIN FUNCTION

The biological function of proteins usually resides in small and evolutionarily conserved units. One popular example is the well-studied catalytic triad of serine proteases [2]; a configuration of only three residues, solely responsible for enzymatic activity. These spatial residue patterns are not necessarily contiguous in protein sequence and usually long-range contacts brought to spatial proximity during the protein folding process (Figure 3.1A). Consequently, it is hard to identify these patterns by multiple sequence alignment (MSA) techniques [102, 103]. The detection and thorough comprehension of structural motifs can help to bridge the gap between protein sequence, structure, and function. Local regions in a protein structure mediate function and their analysis is key in order to understand detailed evolutionary and functional relationships [4, 32]. However, it is yet to be answered, and stated as a major challenge in structural bioinformatics, how naturally or engineered perturbations at molecular level influence the overall protein structure or function [5]. Even smallest changes on crucial residues were shown to shift or disrupt the global fold of proteins (Figure 3.1B) [6]. In conjunction with the ever-increasing number of available macromolecular structures, as discussed in Section 1.1, computational methods for the identification and interpretation of significant spatial residue patterns are experiencing a high demand.

3.2. DEFINITION

Throughout literature the term "structural motif" is not clearly defined. In general, structural motifs described in literature vary in the number of residues between three and six [104, 105]. Furthermore, structural motifs are usually considered to be the first shell [106] residues in direct contact with the substrate of an enzyme. However, this definition does not hold for structural motifs buried in the hydrophobic core of a protein such as structure-stabilizing elements [30]. In the context of this thesis a more general definition of structural motifs is used (Definition 3.1).

Definition 3.1 (Structural motif)**.** Let $T = \{t_1, t_2, \ldots, t_n\} \subset \mathbb{R}^3$ be a protein (a set of amino acids) consisting of n residues. Each residue t is represented by a single point in $\mathbb{R}^3$, e.g. its alpha carbon atom. The set $Q = \{q_1, q_2, \ldots, q_k\} \subset T$ of size k with $k \geq 2$ is called a structural motif. Usually $k \ll n$ and $\forall q_v, q_w \in Q, v \neq w : \|q_v - q_w\| \lessapprox 25$ Å. The residues of a structural motif are in spatial proximity, e.g. part of a ligand binding site.

A structural motif is usually much smaller than the containing protein structure. An illustration of the relative size of a structural motif, present in the Zika virus capsid, is given in Figure 3.2. Furthermore, structural motifs are meant to be general patterns, present in and descriptive for a protein family [107], an enzyme class [2], or a protein superfamily [33]. Hence, structural motifs are evolutionarily conserved to retain protein structure or function (Definition 3.2).

Definition 3.2 (Structural motif conservation)**.** A structural motif $Q \subset T$ observed in a protein T is evolutionarily conserved if there exist other proteins $\mathcal{T}$ similar to T according to at least one the following criteria:

- sequence similarity,
- structure similarity, or
- function similarity (e.g. catalysis of the same chemical reaction).

A conserved structural motif can be observed in each protein $T' \in \mathcal{T}$. Furthermore, all conserved motifs $\mathcal{Q}$ are similar to each other, i.e. $\forall Q_i, Q_j \in \mathcal{Q}, i \neq j : d(Q_i, Q_j) \leq \varepsilon \vee s(Q_i, Q_j) \geq \varepsilon$ with $d(Q_i, Q_j)$ and $s(Q_i, Q_j)$ being a dissimilarity and similarity measure, respectively. The (dis-)similarity cutoff is given by ε.

3.3. BIOLOGICAL ROLES

Structural motifs are the functional determinants for a wide array of cellular processes (see Figure 3.3), for example catalytic activity of enzymes [2], DNA/RNA interaction [109, 110], or ion fixation [111–113]. Additionally, they were observed to aid structure stabilization [30]. Even highly divergent protein superfamilies, such as the Enolase Superfamily (ES), can be represented adequately by structural motif templates [33].

Catalytic Activity The degradation of polypeptides is realized by enzymes which catalyze the cleavage of peptide bonds. This chemical reaction is usually fulfilled by certain substrate specific enzymes, the proteases. The first identified protease catalytic site was revealed in 1967 by studying α-chymotrypsin with X-ray diffraction [114]. Subsequent structure alignments of proteases uncovered a remarkable similarity of the active sites [115]. Consequently, it was found that solely three residues are responsible for peptide bond

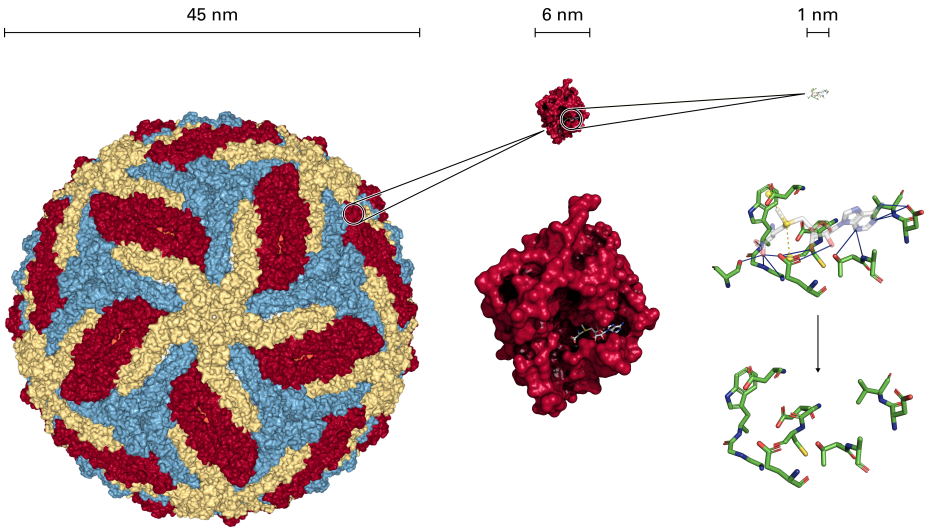


Figure 3.2.: Structural motif of the NS5 methyltransferase (PDB:5kqr), a protein of the Zika virus capsid (PDB:5ire) [108]. The relative size (≈1 nm) as well as the number of atoms (48) is small compared to the whole biological assembly with a size of ≈45 nm and ≈661,000 atoms. Nevertheless, the structural motif is key for host invasion of the virus.

cleavage in proteases, forming a so-called catalytic triad. The triad consists of histidine, aspartic acid, and predominantly serine (Figure 3.3A). For the purpose of conservation and structure stabilization, the catalytic triad is part of an extensive hydrogen bonding network [2]. During the reaction serine attacks the carbonyl group of the peptide bond whereby histidine acts as general base in the first step. The protonated histidine residue is then stabilized by formation of hydrogen bonds with aspartic acid. Finally, the peptide bond is broken by a water-mediated reaction [2].

Structure Stabilization Interactions between aromatic amino acids were shown to be relevant for protein folding [116] or structure stabilization [117]. KOUTSOTOLI AND TZAKOS described the molecular basis of the infection process of human cells with enterohemorrhagic *Escherichia coli*. Here, a double CH-π stacking interaction of two tryptophan residues, sandwiching a proline, is exploited during the infection process. This so-called CH-π interaction motif (Figure 3.3B) was found to be present in more than 600 cases in the PDB [30]. This constitutes an important example to justify the necessity to handle extremely compact structural motifs buried in the core of a protein.

Ion Binding Metal binding sites in proteins are another example where the correct geometric arrangement of residues is essential [118]. The ion coordination center of human ferritin, which is relevant for bioavailable iron storage [112], can be described as a structural motif consisting of five residues. Other examples include the zinc finger binding motif (Figure 3.3C) [119] or the copper coordination center of cupredoxins [107].

Nucleotide Interaction Nucleotide binding proteins require specificity for DNA or RNA recognition sequences to fulfill their regulatory roles. One example of a specific RNA binding motif is shown in Figure 3.3D [110]. This motif specifically binds to the nucleotide sequence YCAY (Y are pyrimidines). The motif's residues supply hydrogen bonds precisely as those that would be normally served by a complementary stand [110].

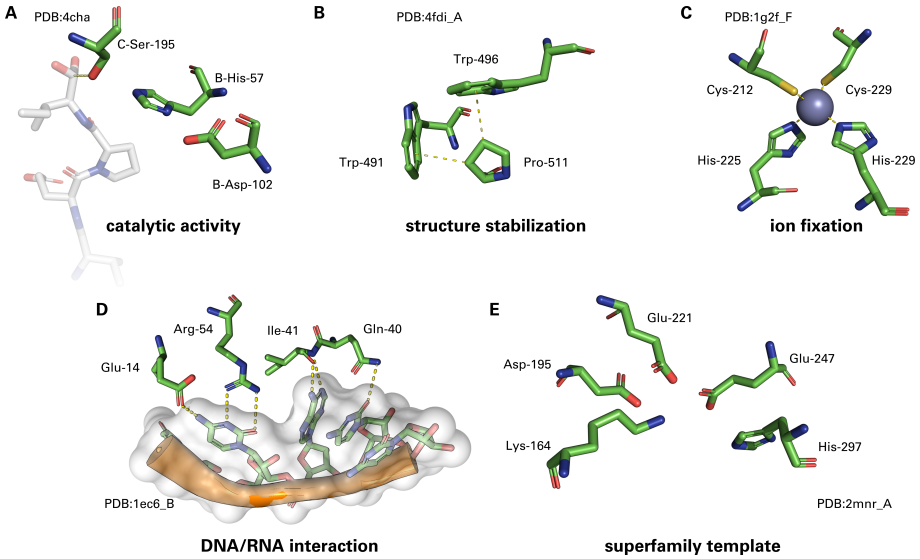


Figure 3.3.: Selected structural motifs and their biological roles. (**A**) The catalytic triad of histidine, aspartic acid, and serine, is responsible for the catalytic activity in serine proteases [2]. (**B**) A double CH-π stacking motif was shown to be a key determinant during the infection process of human cells by enterohemorrhagic *Escherichia coli* [30]. (**C**) The correct coordination of metal ions in protein structures is essential for proper function and realized by ion coordination centers, e.g. the Cys_2His_2 motif of the zinc finger binding domain [119]. (**D**) The specific interaction of nucleotides can require specialized structural motifs, e.g. the YCAY RNA binding motif [110]. (**E**) Structural motifs can be descriptive for whole protein superfamilies, such as the ES template [33].

Superfamily Templates Even if the overall sequence or fold of related proteins diverges during evolution, a common function might be preserved. The ES shares a partial reaction mechanism: the α-proton abstraction of carboxylic acid [120]. A common structural motif has been derived for this superfamily (Figure 3.3E) [33]. Using this template definition it was possible to represent the ES and their subgroups appropriately. Another example of a superfamily-representing motif has been described for the haloacid dehalogenase superfamily [33].

4. COMPUTATIONAL STRUCTURAL MOTIF DETECTION

This chapter incorporates content from the publication "Fit3D: a web application for highly accurate screening of spatial residue patterns in protein structure data" published in *Bioinformatics* as well as from "Unsupervised Discovery of Geometrically Common Structural Motifs and Long-Range Contacts in Protein 3D Structures" published in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. For a detailed list of author contributions please refer to page 17.

In conjunction with the accelerating progress in the experimental determination of protein structures, the computational detection of structural motifs becomes increasingly important. During the last years, structural genomics projects have led to protein structures where the molecular function remains unknown, especially when novel folds are observed [40, 102]. Computational structural motif detection can be applied to bridge the gap between structure and function. Furthermore, detection algorithms are a valuable tool for the detailed characterization of molecular mechanism as shown later for aaRSs.

Motivation The detection of known structural motifs was successfully applied to predict molecular functions [31, 121] or to identify evolutionary relationships [33], which cannot be detected in sequence but represent a key event of convergent evolution [122]. The function of a structure can be predicted by matching against libraries of structural motifs with known function [123], such as the Catalytic Site Atlas (CSA) [124, 125], the Structure-Function Linkage Database (SFLD) [126], or active site patterns [123]. In the following sections the different types of computational structural motif detection are elucidated. The problem of structural motif detection can be divided into two major categories:

I) the template-based and

II) the template-free *de novo* detection of structural motifs.

Furthermore, a summary of existing methods is given. Based on the features and properties of these approaches, limitations are concluded which have to be addressed in this thesis.

4.1. TEMPLATE-BASED

Template-based structural motif detection has been successfully used, for example, to predict the function of proteins [121] as elucidated beforehand. Hence, there is a plethora of methods available that are capable of detecting matches of given template structural motifs guided by a dissimilarity measure or score. Commonly, the root-mean-squared deviation (RMSD) of atom coordinates after superimposition is used, calculated according to the equation

$$RMSD(A, B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|a_i - b_i|} \tag{4.1}$$

with $A = \{a_1, a_2, ..., a_i\}$ and $B = \{b_1, b_2, ..., b_i\}$ being ordered sets of atoms of the same size. The ordering of the two sets defines which atoms should be paired, e.g. a_1 is paired with b_1. The RMSD is a dissimilarity measure.

Some template-based methods use graph-based concepts [31, 127, 128], others are based on distance comparison [129, 130]. However, the field of application remains common: there is one structural motif that is known *a priori* and should be found in one or more target structures to infer a common function, fold, or ancestry. Template-based structural motif detection is defined in Definition 4.1 and an illustration is given in Figure 4.1.

Definition 4.1 (Template-based structural motif detection)**.** Let $\mathcal{T}$ be a set of protein structures, where a given template structural motif $\hat{Q}$ should be detected. Template-based structural motif detection employs $\hat{Q}$ as template to detect motifs $\mathcal{Q} = \{Q_1, Q_2, ..., Q_n\}$ similar to $\hat{Q}$ according to a (dis-)similarity measure: Hence, $\forall Q_i \in \mathcal{Q} : d(\hat{Q}, Q_i) \leq \varepsilon \vee s(\hat{Q}, Q_i) \geq \varepsilon$. The (dis-)similarity cutoff is given by ε.

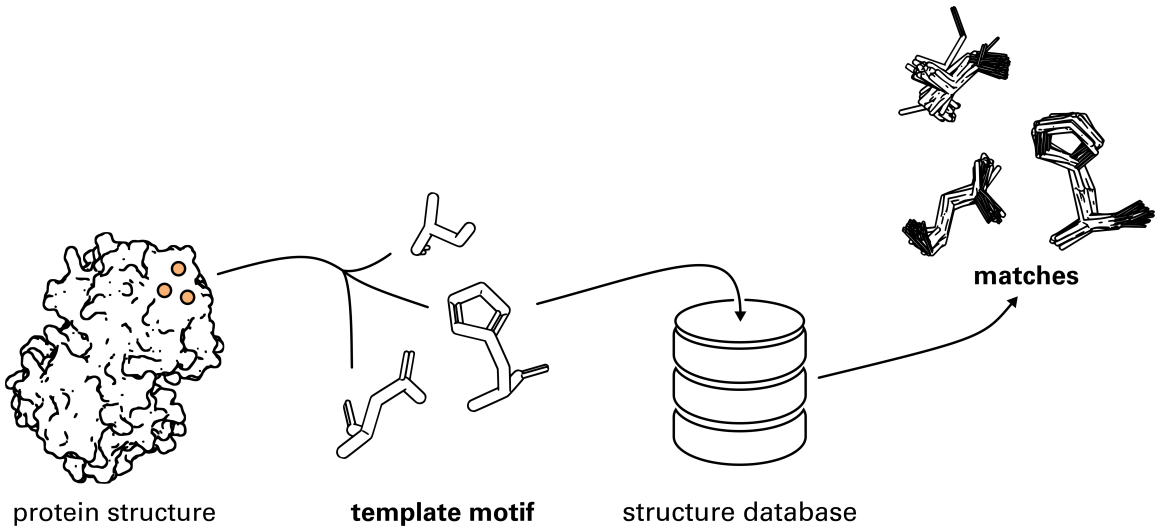


Figure 4.1.: The process of template-based structural motif detection. A single protein serves as starting point to define a template motif, e.g. consisting of residues with known catalytic activity (orange circles). Subsequently, the template motif is used to detect similar matches in sets of protein structures derived from databases such as the PDB [38]. These matches can be used to infer common function or ancestry.

4.2. TEMPLATE-FREE

The template-free and unsupervised detection of structural motifs without any *a priori* knowledge remains challenging. In contrast to template-based structural motif detection, the task is to find residue patterns which are shared across a given dataset of protein structures. The dataset may consist of a pair of proteins, e.g. to identify common binding sites, or hundreds to thousands of protein structures. Template-free structural motif detection eliminates the need for a template motif, which is often times unknown. Algorithms addressing this problem are usually related to or based on data mining techniques, such as graph mining [131] and pattern recognition [104]. The results obtained by template-free approaches can in turn be used to define structural motifs for template-based methods. Figure 4.2 illustrates the workflow for template-free structural motif detection. A more formal definition of template-free structural motif detection is given in Definition 4.2.

Definition 4.2 (Template-free structural motif detection)**.** Let $\mathcal{T}$ be a set of protein structures with shared properties such as common function or conserved fold. Template-free structural motif detection is employed to discover a set of similar and recurrent structural motifs $\mathcal{Q} = \{Q_1, Q_2, ..., Q_n\}$, such that $\forall Q_i \in \mathcal{Q}\ \exists\ T \in \mathcal{T} : Q_i \subset T$. Additionally, all structural motifs in $\mathcal{Q}$ must be similar to each other according to a (dis-)similarity measure: $\forall Q_{i,j} \in \mathcal{Q} : d(Q_i, Q_j) \leq \varepsilon \vee s(Q_i, Q_j) \geq \varepsilon$. The (dis-)similarity cutoff is given by ε.

4.3. AVAILABLE SOFTWARE

Both problems – template-based and template-free detection – were subject of many studies with emphasis on template-based detection. Hence, algorithms and software tools exist, which address both problems. Especially for template-based detection many methods were developed and thus the selection of the presented methods is not exhaustive, but aims at covering the most important tools. The methods are classified according to their underlying computational strategy, which usually derives from classical computer science problems. An overview of the assessed methods is shown in Table 4.1 and Table 4.2. Additional summaries of methods can be found in references [32, 133, 134].

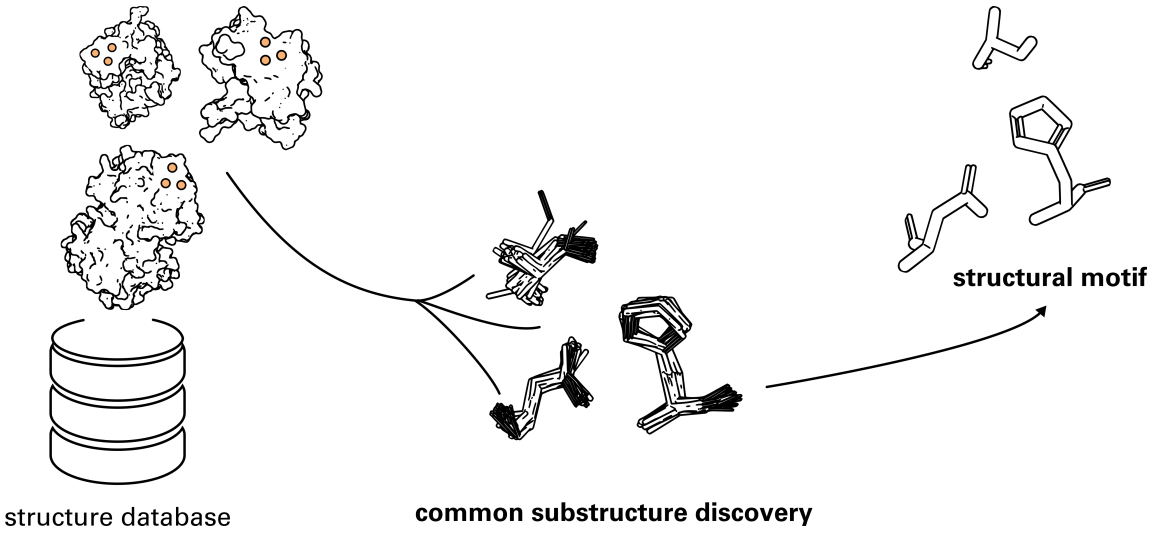


Figure 4.2.: The general workflow for the template-free detection of structural motifs. A set of proteins with presumably shared function or other common properties is derived from structure databases, e.g. the Pfam database [132]. By applying algorithms for common substructure discovery, structural motifs can be identified, which are recurrent in the initial dataset (orange circles). These patterns are likely to be specific for the initial protein family, important for protein structure, or function.

4.3.1. ASSESSED FEATURES

The following review of existing methods for structural motif detection focuses on different aspects of the methods. A set of desirable features was composed; required for general-purpose methods. Most importantly, these include flexibility in the computational representation of structural motifs. Because most structural motif detection approaches include a step to assess the geometric similarity of matches, usually by determining the RMSD after superimposition of atoms, the adequate algorithmic representation of residues is a critical step. Another relevant feature is the detection of matches that are located in multiple protein chains (so-called inter-molecular matches), which are especially important to locate structural motifs at protein-protein interfaces. The possibility to define constraints such that alternative residues are allowed to be matched at defined positions of the template motif, so-called position-specific exchanges (PSEs), is another desirable feature if sequence conservation of structural motif residues is not given.

Furthermore, general-purpose methods should feature an assessment of the statistical significance of reported matches and should allow the definition of mixed structural motifs, e.g. matching of DNA/RNA, or motifs which contain ligands or ions. Ideally, implementations of the method should be provided that are easy-to-use for the non-expert user or allow the integration into custom processing pipelines.

4.3.2. TEMPLATE-BASED

In the following, template-based structural motif detection approaches are presented that can be grouped roughly in five major categories of underlying computer science problems: combinatorial, dynamic programming, geometric hashing, graph-based, and tree search. Combinatorial approaches use strategies such as the intelligent testing of permutations of residues to detect structural motif matches. Dynamic programming is used to determine the optimal sequential alignments of match candidates before superimposition. Methods based on geometric hashing, originally suggested for computer vision [135], usually require the preprocessing of the structure database before the actual search. Graph-theoretical approaches abstract protein structures to graphs and use graph comparison algorithms to

identify potential matches, whereas tree search algorithms can be applied for the iterative identification of matches. The assessment of geometric similarity between candidates and the template motif is usually based on the superimposition of atom coordinates, e.g. by the Kabsch algorithm [136], and the subsequent determination of the RMSD.

Combinatorial He et al. published BALLAST, a motif matching algorithm that supports the representations of protein residues by their alpha carbon, beta carbon, or side chain centroid coordinates. BALLAST is based on the comparison of residue composition and geometry in local environments [137]. These local environments are constructed by extracting all residues within a ball of defined radius around the currently considered residue. In contrast, Query3D is based on the evaluation of different criteria to reduce the combinatorial complexity: residues of the template motif must be neighbored in the target structure, geometric and biochemical similarity must be given [138].

Dynamic Programming The APoc algorithm uses dynamic programming to determine the optimal sequential alignment of candidate residues against the template structural motif [139]. Initially, guessed solutions are generated, which are subsequently used as a starting point to determine the optimal sequential alignment via dynamic programming. pvSOAR [140] relies on the comparison of surface patterns, which are encoded in sequence fragments, and compared by dynamic programming [141].

Geometric Hashing Moll et al. presented the LabelHash algorithm for detecting structural motifs in proteins [142]. At first, a hash table is constructed for the database of target proteins which includes *n*-tuples of residues that satisfy certain geometric constraints. Hence, the preliminary construction of an individual hash table is necessary for each set of target structures [143]. Secondly, a hashed representation of the template motif is looked up in this table, which allows the retrieval of geometrically similar matches. SiteEngine [144] operates similarly but exclusively considers the protein surface.

Graph-Based ProBiS represents the protein surface residues as vertices of a graph [127, 145]. The maximum clique algorithm is employed to discover agreements between the template and target graph. Matches between these graphs are equivalent to local structure similarities on the protein surface [127]. ASSAM [128] can be assigned to the same algorithmic category. It represents the protein structure as a graph, where individual side chains of amino acids are vertices with edges describing their geometric relationships. Each vertex in the graph characterizes two pseudoatoms, which in turn represent the functional part of the side chains. An algorithm for subgraph isomorphism is used to identify structural relationships between the template motif and target structures [128]. Nilmeier et al. developed the CatSId method to identify matches of structural motif templates, derived from the CSA, in a set of target structures. Template motif and target structure are converted to distance matrices and subsequently graphs are constructed to identify matches by subgraph isomorphism detection [31].

Tree Search RASMOT-3D PRO relies on the representation of residues by their alpha and beta carbon atoms exclusively. The matching of a template motif is based on the comparison of inter-atomic distances and the subsequent calculation of the RMSD [130]. Although not explicitly stated by the authors, this corresponds to the iterative construction of a search tree to find suitable candidates. The SPASM algorithm [129] operates similarly and uses a depth-first search algorithm as well as additional constraints to prune the search

tree as early as possible. SPASM allows to represent the template motif by alpha carbon atoms and the centers of gravity of side chain atoms [129].

Other Other methods, which cannot be attributed to one of the former categories, are CMASA [146], eMatchSite [44], GASS [43], and Suns [47]. CMASA detects local structure similarity through a contact matrix average deviation technique [146]. eMatchSite uses machine learning to estimate all-against-all alpha carbon distances and the Kuhn-Munkres algorithm to find alignments between binding site residues [44]. The GASS algorithm is based on the intelligent heuristic application of a genetic algorithm to identify matches of template structural motifs in a nondeterministic way [43]. Suns is available as plug-in for the PyMOL software [147] and allows real-time structural motif detection. Suns follows the idea of a web search engine and divides protein structures into "pages" (volumes) and "words" (chemical motifs). Indexing strategies and refinement steps are used to optimize queries [47].

4.3.3. TEMPLATE-FREE

In contrast to template-based methods, the template-free and unsupervised detection of structural motifs without any *a priori* knowledge is barely addressed. The existing algorithmic approaches can be divided into three major categories: itemset mining, graph-based, and string matching. Itemset mining originated from data mining ("market basket analysis") and was used for different applications, such as text mining, time-series, graph, or spatial data analysis [148]. Graph-based methods represent protein structures as graphs and use, for example, frequent subgraph mining to identify similar structural motifs in the dataset. Algorithms based on string matching encode structural features into strings, which are then matched to identify potential correspondences between the structures of origin.

Itemset Mining The generalized aim of itemset mining is to determine associations between items, which can be expressed by different measurements (metrics). One basic metric is the support – a measurement that describes the relative occurrence of sets of items ("itemsets") in the data. Itemset mining was already applied successfully to protein data, e.g. to identify binding motifs for transcription factors or splicing patterns [149]. However, in 2014 ZHOU ET AL. were the first to suggest using frequent itemset mining to spot interesting biological patterns in protein structure data without any abstraction of spatial data to graphs, distance matrices, or structural features [150]. They introduced the concept of cohesion to avoid the explicit restriction of distances between items during the mining process and to discover patterns in spatial proximity, i.e. cohesive patterns. The method was applied to different protein families and revealed cohesive patterns that span large distances at sequence level but are brought to proximity in tertiary structure by protein folding. The authors suggest that these patterns play a role for the specific or overall structure of the protein [150]. In a subsequent study authored by the same group, their methods were extensively applied to a dataset representative for the whole PDB in order to identify cohesive patterns not linked to any concrete fold or function. These patterns occur mainly beyond annotated Pfam domains [151]. Additionally, the method was used to mine specific cohesive patterns, which correlate with optimal growth temperature of different prokaryotic species, and to identify preferential contacts in DNA-binding proteins [151].

Graph-Based For all graph-based methods the protein structures have to be converted to graph representations in the first instance, which are then mined for frequent subgraphs. The method of DHIFLI ET AL. reduces the obtained data from graph mining by considering

metadata, such as evolutionary information for similarities of amino acids derived from substitution matrices [152]. Frequently occurring residue packing patterns, mined as frequent subgraphs, were successfully exploited to determine protein family association [153]. In 2008, XIE AND BOURNE presented a method to identify local similarities of unknown binding sites between proteins, based on sequence profile alignments by incorporating evolutionary information and representing the protein structure as a graph [154]. This concept was further extended with sophisticated scoring schemes that incorporate not only spatial similarity, but also physicochemical and other biologically relevant features [44].

String Matching The approach of DUDEV AND LIM converts fragments of proteins structures to strings using a structural alphabet [104]. These strings are then compared to each other in order to identify structural motifs which bind magnesium ions. Moreover, the mining of sequence patterns using string matching was successfully applied to identify packing patterns in diverse protein families [103, 105]. The SPratt2 [105] algorithm uses a structural alphabet to represent proximal spatial patterns and includes secondary structure information.

Other BCSearch uses an innovative alignment-free and transformation-invariant method, the so-called Binet-Cauchy kernel [155], to determine local similarities in a set of protein structures [156]. However, BCSearch is limited to the detection of structural motifs that are consecutive in the protein sequence. Methods that rely on plain spatial data without abstraction to graphs or structural alphabets include FunClust [157] and LGA [158]. These methods are designed to be used with a small set of protein structures, 20 in the case of FunClust, or a pair of protein structures for LGA. FunClust uses the Query3D algorithm [138], which evaluates different criteria to find a set of common residues between pairs of proteins by combinatorial extension. LGA operates similarly and tries to find the longest segments of residues that fit under a specified RMSD cutoff [158].

4.4. LIMITATIONS

Although a variety of approaches for structural motif detection have been presented that involve local residue comparison, none of them incorporate a geometric evaluation of match candidates at the greatest possible detail. Due to the available high-resolution structure data, algorithms should be capable to detect similarities at the atom level in order to identify similar structural properties, such as conserved side chain orientations or confined backbone traces.

Computational Representation Most of the presented algorithms are limited in terms of the computational representation of structural motifs. Template-based approaches usually rely on a reduced representation of motif residues, e.g. by their alpha or beta carbon atoms [130] or side chain centroids [129]. This can be especially crucial for the detection of highly specific ligand interactions that are mediated by residue side chains, which is, as shown later, the case for Class II aaRSs. Hence, a strong limitation of existing methodologies is how structural motifs are represented algorithmically. Nearly all presented template-free approaches rely on the abstraction of protein structures as graphs [131] or structural alphabets [104] and are thus even more limited to detect similarities at atom level.

Isofunctional Mutations Another important point is the consideration of isofunctional mutations in structural motifs, which are regular events during protein evolution and mandatory for the robustness and evolvability of binding site residues [4]. This can include, for

Table 4.1.: An overview of template-based structural motif detection methods grouped by the underlying algorithmic concept: combinatorial (CO), dynamic programming (DP), geometric hashing (GH), graph-based (GB), tree search (TS), or other (OT). The number of citations of each method is indicated by dots: • <10, •• <100, ••• >100 (taken from scholar.google.de).

			features						use		
	name	reference	custom atom representation	inter-molecular motifs	PSEs	statistical significance	DNA/RNA motifs	ligand motifs	implementation	open source	limitation
CO	BALLAST	[137] •	-	-	-	-	-	-	-	-	
	Query3d	[138] ••	-	-	-	-	-	-	✓	✓	
DP	APoc	[139] ••	-	✓	-	✓	-	-	✓	✓	ligand binding pockets
	pvSOAR	[140] ••	-	-	-	✓	-	-	✓	-	protein surface
GH	LabelHash	[143] •	-	✓	✓	✓	-	-	✓	-	
	SiteEngine	[144] •••	-	✓	-	-	-	-	✓	-	protein surface
GB	ASSAM	[128] ••	-	✓	-	-	-	-	✓	-	
	CatSId	[31] ••	-	-	✓	-	-	✓	✓	-	
	ProBIS	[127] •••	-	✓	-	✓	-	-	✓	✓	protein surface
TS	RASMOT-3D PRO	[130] ••	-	-	-	-	-	-	✓	-	results limited
	SPASM	[129] •••	-	✓	✓	-	-	-	✓	-	deprecated search database (2008)
OT	CMASA	[146] ••	-	✓	-	✓	-	-	✓	✓	
	eMatchSite	[44] ••	-	-	-	-	-	✓	✓	✓	ligand binding pockets
	GASS	[43] •	-	✓	-	✓	-	-	✓	✓	
	Suns	[47] •	✓	✓	-	-	-	-	✓	✓	PyMOL plug-in

Table 4.2.: An overview of template-free structural motif detection methods grouped by the underlying algorithmic concept: itemset mining (IM), graph-based (GB), string matching (SM), or other (OT). The number of citations of each method is indicated by dots: • <10, •• <100, ••• >100 (taken form scholar.google.de).

			features				use		
	name	reference	custom atom representation	statistical significance	DNA/RNA motifs	ligand motifs	implementation	open source	limitation
IM	FreSCOs	[151]••	-	✓	✓	-	-	-	set definition
GB	Dhifli et al.	[152]••	-	-	-	-	-	-	representative pattern selection
	Huan et al.	[131]••	-	✓	-	-	-	-	
	SOIPPA	[154]•••	-	✓	-	-	-	-	pairs of proteins
SM	Dudev and Lim	[104]••	-	✓	-	-	-	-	
	SPratt2	[105]••	-	-	-	-	-	-	
OT	BCSearch	[156]•	-	✓	-	-	✓	-	consecutive fragments
	FunClust	[157]••	-	✓	-	-	✓	-	maximal 20 protein structures
	LGA	[158]•••	-	✓	-	-	✓	-	pairs of proteins

example, the switching of a proton donor from lysine to histidine. Current methods for template-based structural motif detection lack the possibility to specify such search constraints. Hence, the consideration of residue exchanges for each position of the template motif is desirable to mimic residue substitutions, which might have occurred during protein evolution or were induced by directed mutagenesis or protein design.

Field of Application There are several structural motif detection algorithms that are tailored to a specific application, e.g. ProBIS [127] or pvSOAR [140] for the detection of structural motifs on the protein surface. However, this can be a critical limitation if, for example, structural motifs should be considered that are buried in the hydrophobic core of the protein and are relevant for intrinsic structure stabilization [30]. Another problem is evident for some presented template-free structural motif detection algorithms; they are restricted to a few protein structures in the target dataset such as FunClust [157] or LGA [158]. This is insufficient if common structural motifs should be found in families of proteins where a larger set of protein structures is available.

Usability A common problem of structural motif detection algorithms is the availability of implementations and up-to-date search databases as well as usability. The authors of most template-based approaches provide implementations of their algorithms, e.g. LabelHash [143] or ProBIS [127]. However, frequently there are some limitations in terms of usability such as for the RASMOT-3D PRO web server [130], where only a limited number of matches is reported. The availability of an open source implementation or an application programming interface (API) is a strong plus for advanced users and allows for the flexible integration into custom processing pipelines. Eight of fifteen template-based approaches do not meet this requirement. Only three out of nine methods for template-free structural motif detection provide an implementation, while no open source implementations are available at all.

General-Purpose Tools In general, there is a lack of versatile and general-purpose algorithms for the template-based and template-free detection of structural motifs. The main drawbacks (Figure 4.3) of current methods that should be addressed in this thesis are:

- the computational representation of structural motifs by an arbitrary selection of atoms without prior abstraction and,
- the consideration of residue exchanges for each position of the template motif, so-called PSEs.

Especially, the former is of high importance to benefit from contemporary structure data of atomic resolution. Moreover, the support to detect inter-molecular occurrences of structural motifs lacks for most tools. However, structural motifs which occur in multiple protein chains have been shown to be of functional importance [30, 112, 159]. Furthermore, structural motifs in DNA or RNA structures should be supported to promote the application of structural motif detection algorithms for these macromolecules, e.g. for the structural study of riboswitches [160]. Although some scoring schemes for structural motif similarity were presented [161, 162], most methods do not report the statistical significance of matches, which makes it difficult to assess their relevance. This thesis tackles the mentioned limitations and aims at the combination and extension of advantages of existing methods to enable the versatile analysis of structural motifs. The developed algorithms are applied to aaRS enzymes to study their molecular recognition mechanisms at atomic resolution.

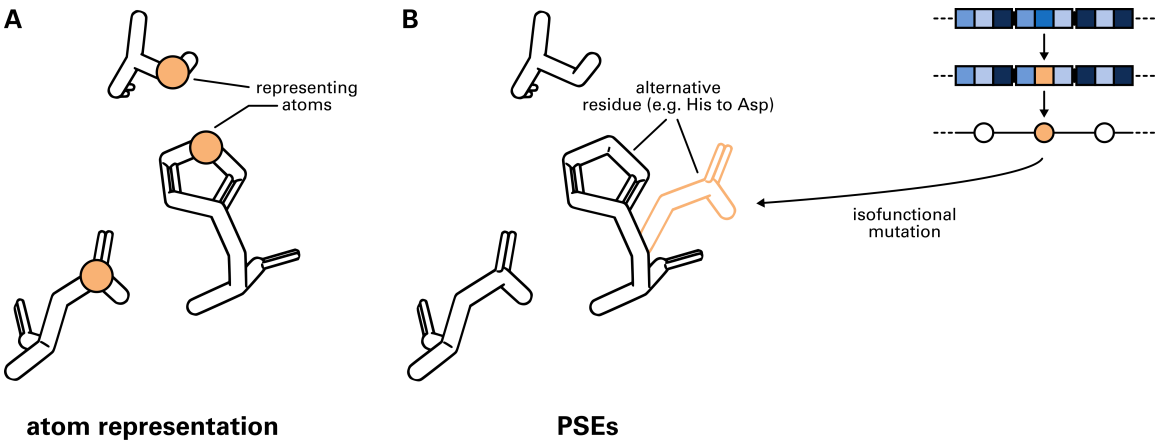


Figure 4.3.: The main limitations of structural motif detection algorithms. (**A**) The computational representation of structural motifs determines the sensitivity and specificity of the method. (**B**) The definition of PSEs allows detecting occurrences of structural motifs which underwent isofunctional mutations during protein evolution.

Part II.

RESULTS

5. STRUCTURAL MOTIFS IN AMINOACYL-TRNA SYNTHETASES

This chapter is based on the results of the article "Backbone Brackets and Arginine Tweezers delineate Class I and Class II Aminoacyl tRNA Synthetases" published in *PLOS Computational Biology*. For a detailed list of author contributions please refer to page 17.

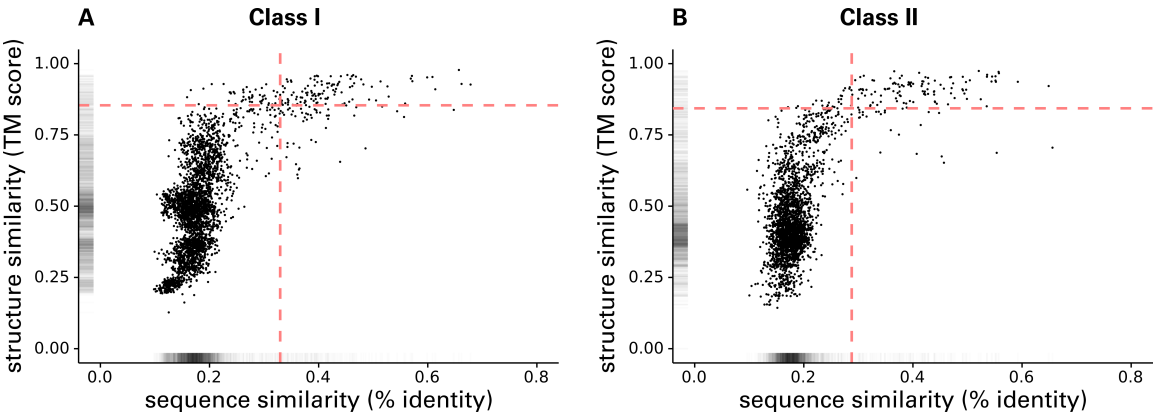


Figure 5.1.: Pairwise sequence and structure similarity of non-redundant cluster representative chains for Class I (**A**) and Class II (**B**) aaRSs. Depicted is the sequence similarity (% identity) after a global Needleman-Wunsch [165] alignment of both structures against the structure similarity determined with by TM-align [164]. For Class I (Class II) 95% of all pairs exhibit <33% (29%) sequence identity and <0.85 (0.84) TM score. The 95% quantile borders are depicted as red dashed lines.

The following chapter presents the results of structural motif detection in aaRSs. The creation of a manually curated dataset of almost 1,000 aaRS structures, presented in [163], constitutes the basis for all analyses. A thorough structural characterization of the motifs in Class I and Class II aaRSs is complemented by template-based and template-free detection results.

5.1. DATASET OF STRUCTURES

In order to identify structural motifs in aaRSs, relevant for the molecular recognition mechanism, a dataset was generated that contains ligand information. It is composed of 972 individual chains containing 448 (524) Class I (Class II) catalytic aaRS domains and covers at least one ligand-bound structure for each aaRS Type. The dataset is provided as Supporting Information of the article "Backbone Brackets and Arginine Tweezers delineate Class I and Class II Aminoacyl tRNA Synthetases" by KAISER ET AL. [163].

The pairwise sequence identity of the dataset is below 33% (29%) for 95% of all Class I (Class II) structures, while pairwise structure similarity is high with a TM score [164] over 0.8 for 95% of the structures (Figure 5.1). The high sequence diversity probably stems from the variety of covered organisms and domain insertions. In contrast, the low structure diversity can be seen as a result of conserved function and the shared topology of the catalytic domain within each aaRS Class.

Sequence positions of all structures in the dataset were unified using an MSA generated with the T-Coffee expresso pipeline [166] (see Methods in [163]). This type of MSA is backed by the additional structural alignment of protein structures. Hence, the structurally conserved catalytic core region is preferred during alignment, since domain insertions and attachments do not align structurally across the whole dataset. The MSA allows the investigation of a plethora of structures independently of the concrete aaRS Type. This investigation is aided by a renumeration that effectively provides a means to compare sequentially divergent, structurally similar proteins. All further referenced positions are given in accordance to this MSA. In figures where depictions of structures are shown, the original sequence positions of residues are listed. To infer original sequence positions from given renumbered sequence positions, mapping tables are provided as Supporting Information alongside the published article [163]. These tables contain the corresponding original sequence positions for each position of the MSA and for each structure in the dataset.

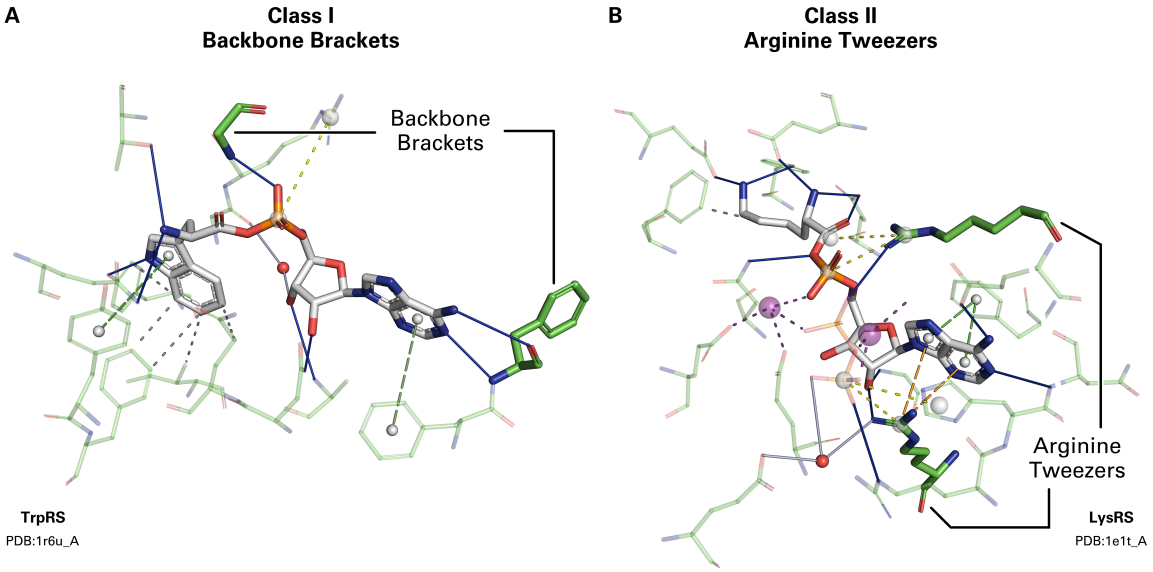


Figure 5.2.: Both aaRS classes contain highly conserved patterns, responsible for proper binding of the ATP ligand. Class I structures share a conserved set of backbone hydrogen interactions with the ligand: the Backbone Brackets. Class II active sites contain a pattern of two arginine residues grasping the ATP ligand: the Arginine Tweezers. Interactions were calculated with PLIP [167] and are represented with colored (dashed) lines: hydrogen bonds (solid, blue), π-stacking interactions (dashed, green), π-cation interactions (dashed, orange), salt bridges (dashed, yellow), metal complexes (dashed, purple), and hydrophobic contacts (dashed gray). (**A**) Class I Backbone Brackets motif and interactions with the ligand Tryptophanyl-5'AMP as observed in tryptophanyl-tRNA synthetase (TrpRS) structure PDB:1r6u chain A. (**B**) Class II Arginine Tweezers motif and interactions with the ligand Lysyl-5'AMP as observed in lysyl-tRNA synthetase (LysRS) structure PDB:1e1t chain A.

In order to investigate the contacts between aaRS residues and their ligands, noncovalent protein-ligand interactions were annotated with PLIP [167]. For simplification, "ATP ligand" refers to all forms of ligands in aaRS binding mode M1 (see Section 2.5) that contain an adenosine phosphate substructure including aminoacyl adenylate and AMP. Manual investigation of these contacts revealed two highly consistent interaction patterns between catalytic site residues and the ATP ligand: conserved backbone hydrogen bonds in Class I as well as two arginine residues with conserved salt bridges and side chain orientations in Class II. These interactions are part of the extensive noncovalent interaction network in the binding site of aaRSs shown in Figure 5.2. For an overview of the different noncovalent interactions that can be observed between ligands and proteins please refer to [168].

5.2. BACKBONE HYDROGEN BONDS: BACKBONE BRACKETS

Strikingly, the residues mediating the backbone interactions were mapped in 441 of 448 (98%) Class I renumbered structures at the two positions 274 and 1361. Closer investigation at the structural level revealed geometrically highly conserved hydrogen bonds between the peptide bond nitrogen or oxygen atom and the adenosine phosphate part of the ligand (Figure 5.3A). These two residues mimic a bracket-like geometry (Figure 5.3B), enclosing the adenosine phosphate, and were thus termed Backbone Brackets. The interacting amino acids are not limited to specific residues as their side chains do not form any ligand contacts. Hence, position 274 of the Class I motif is not apparent at sequence level while position 1361 exhibits preference for hydrophobic amino acids, e.g. leucine, valine, or isoleucine (Figure 5.3C). Examples for the Backbone Brackets motif are residues 153 (corresponding to renumbered residue 274) and 405 (corresponding to renumbered residue 1361) in Class I ArgRS structure PDB:1f7u chain A.

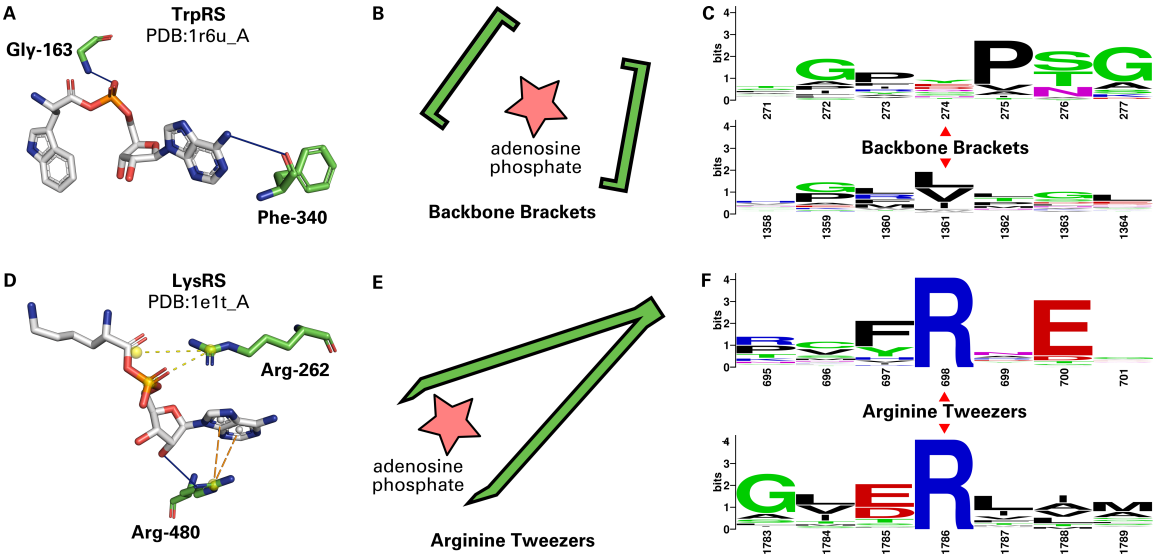


Figure 5.3.: A comparison of the Backbone Brackets and Arginine Tweezers. (**A**) Structural representation of the Backbone Brackets motif interacting with Tryptophanyl-5'AMP ligand in TrpRS (PDB:1r6u chain A). The ligand interaction is mediated by backbone hydrogen bonds (solid blue lines). Residue numbers are given in accordance to the structure of origin. (**B**) The geometry of the Backbone Brackets motif resembles brackets encircling the ligand. (**C**) WebLogo [169] representation of the sequence of Backbone Brackets residues (274 and 1361) and three surrounding sequence positions. Residue numbers are given in accordance to the MSA. (**D**) Structural representation of the Arginine Tweezers motif in interaction with Lysyl-5'AMP ligand in LysRS (PDB:1e1t chain A). Salt bridges (yellow dashed lines) as well as π-cation interactions (orange dashed lines) are established. Residue numbers are given in accordance to the structure of origin. (**E**) The Arginine Tweezers geometry mimics a pair of tweezers grasping the ligand. (**F**) Sequence of Arginine Tweezers residues (698 and 1786) and surrounding sequence positions. The Backbone Brackets show nearly no conservation at sequence level since backbone interactions can be established by all amino acids, while the Arginine Tweezers rely on salt bridge interactions, always mediated by two arginine residues. Residue numbers are given in accordance to the MSA.

5.3. SIDE CHAIN INTERACTIONS: ARGININE TWEEZERS

In contrast, Class II aaRS structures show a conserved interaction pattern of two arginine residues at renumbered positions 698 and 1786, which were identified in 482 of 524 (92%) structures. The two arginine residues grasp the adenosine phosphate part of the ligand (Figure 5.3D) with their side chains, resembling a pair of tweezers (Figure 5.3E), and were thus named Arginine Tweezers. These two residues are invariant in sequence (Figure 5.3F). Examples for the Arginine Tweezers motif are residues 217 (corresponding to renumbered residue 698) and 537 (corresponding to renumbered residue 1786) in Class II AspRS structure PDB:1c0a chain A. Additionally, a highly conserved glutamic acid is the most prevalent amino acid at renumbered position 700. This residue establishes hydrogen bonds to the adenine group of the ligand in seryl-tRNA synthetase (SerRS), HisRS, threonyl-tRNA synthetase (ThrRS), LysRS, prolyl-tRNA synthetase (ProRS), and AspRS.

5.4. APPLICATION OF FIT3D

The flexibility of the structural motif detection algorithms, presented in thesis Chapter 6, allowed a comprehensive structural characterization of the Backbone Brackets and Arginine Tweezers motifs. The following sections describe the structural properties of both motifs, present the results of a PDB-wide screening for similar occurrences, and highlight geometrically conserved regions in the catalytic core domain of aaRSs.

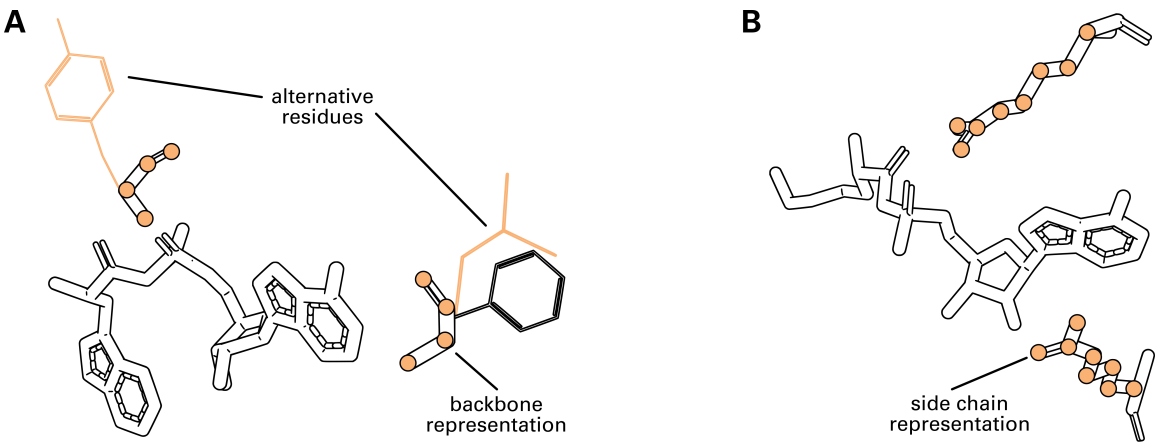


Figure 5.4.: The special requirements for the structural motif detection of Backbone Brackets and Arginine Tweezers. (**A**) The Backbone Brackets motif requires the computational representation via its backbone atoms to maximize sensitivity and specificity during structural motif detection. Due to the low sequence conservation, PSEs have to be defined in order to allow the detection of matches consisting of alternative residues. (**B**) Due to its highly specific interaction with the ATP ligand, the Arginine Tweezers motif can only be reliably detected if side chain atoms of the residues are considered during matching.

Applicability The structural motifs in aaRSs are a perfect example to demonstrate the previously described requirements for general-purpose algorithms for structural motif detection (Section 4.4). The unique properties of the Backbone Brackets and the Arginine Tweezers require specialized features for the applied structural motif detection algorithms: the computational representation based on a defined set of atoms and the definition of PSEs (Figure 5.4). In order to detect the Backbone Brackets the definition of PSEs is essential. Due to the variability in protein sequence (see Figure 5.3C), the consideration of alternative residues at both positions of the motif is necessary (Figure 5.4A). Furthermore, the computational representation should include all backbone atoms to maximize sensitivity and avoid the detection of unspecific matches, which occur with increased frequency if only alpha carbon atoms are considered. Side chain atoms have to be ignored for the Backbone Brackets. The Arginine Tweezers motif features a highly specific interaction with the ATP ligand for which the correct orientation of the side chains is key. Hence, a single-point representation of the Arginine Tweezers is rendered unfeasible; an adequate computational representation must include the side chain atoms of arginine (Figure 5.4B).

5.4.1. STRUCTURAL CHARACTERIZATION

The following section presents the results of an extensive structural characterization of the Backbone Brackets and the Arginine Tweezers. Most of the analyses were performed with the API version of Fit3D [170] in order to implement the specific requirements.

Structural Motif Alignments Prior to a detailed geometric analysis, both structural motifs were aligned in respect to their binding modes M1 and M2 (see Figure 2.5 for a visual representation of M1 and M2) using Fit3D [133]. The alignments (Figure 5.5) were computed based on the backbone atoms of the motifs. The obtained alignments visually support the differences in side chain orientation, alpha carbon distance, and the variable amino acid composition of the Backbone Brackets. Here, the flexibility of the Fit3D algorithm was utilized to quantify the structural variation of the amino acid side chains and backbones, respectively. In general, a high structural conservation of the backbone atoms in respect to

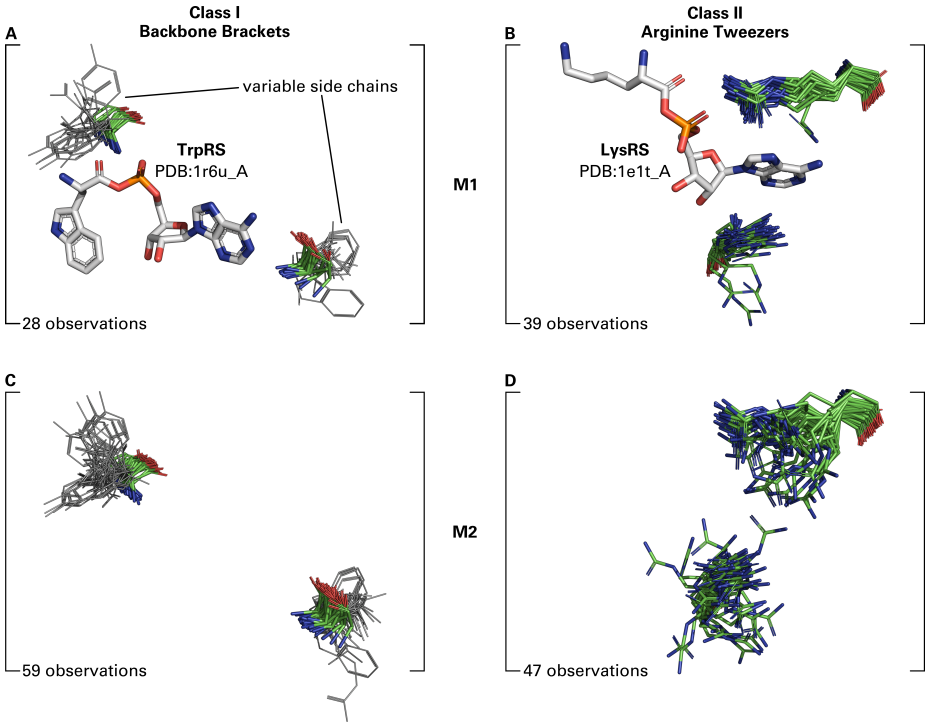


Figure 5.5.: Structural backbone-only alignments of relevant binding site motifs computed with Fit3D [133]. Alignments are grouped by structures derived from ATP bound (M1) and unbound state (M2) for Class I and Class II aaRSs. (**A**,**C**) The Class I Backbone Brackets motif aligned in respect to M1 and M2. A high side chain variance (gray line representation) is evident if an ATP ligand (exemplarily taken from TrpRS, PDB:1r6u chain A) is bound (**A**) and if the ligand is absent (**C**). However, backbone orientations are highly conserved in order to realize consistent hydrogen bond interactions with the ATP ligand. (**B**,**D**) The Class II Arginine Tweezers motif aligned in respect M1 and M2. Low side chain variance can be observed if an ATP ligand (exemplarily taken from LysRS, PDB:1e1t chain A) is bound (**B**), whereas the absence of an ATP ligand (**D**) allows an increased degree of freedom for side chain movement.

the binding modes can be observed for both structural motifs, with a difference in RMSD between M1 and M2 of 0.02 Å for Backbone Brackets and 0.04 Å for Arginine Tweezers, respectively. However, the average side chain RMSD of the Backbone Brackets is continuously high with 1.42 Å, independently of whether ATP is bound or not. This does not hold for the Arginine Tweezers motifs, where the side chain RMSD is low if ATP is bound (M1, 0.72 Å) due to the side chain mediated interactions. In contrast, the side chain variance is high if ATP is not bound (M2, 1.38 Å). This sums up to a change in RMSD of 0.66 Å. Averaged values of backbone and side chain RMSD are listed in Table 5.1.

Geometric Analysis Backbone Brackets and Arginine Tweezers were analyzed at the geometric level to further substantiate the profound differences in ATP recognition. After alignment, the occurrences of Backbone Brackets and Arginine Tweezers were analyzed regarding the distance between their alpha carbon atoms and the relative orientation of their side chains (Figure 5.6). The side chains of the Backbone Brackets residues are expected to exhibit higher degrees of freedom in comparison to the Arginine Tweezers. Furthermore, a significant change in alpha carbon distance of both motif residues indicates a conformational change during ligand binding. The state complexed with ATP (M1) and the state in which no ATP is bound (M2) were analyzed separately in order to quantify these aspects. The angle between side chains of the Backbone Brackets is continuously high: a mean of 144.90 ± 20.93° for M1 and 141.40 ± 20.13° for M2, respectively. This emphasizes that the side chain orientation is indistinguishable between M1 and M2 as only the backbone participates in ligand binding. The alpha carbon distance is conserved for the majority of the

Table 5.1.: Averaged backbone and side chain RMSD values of Backbone Brackets and Arginine Tweezers after superimposition in respect to binding modes M1 and M2. Averaged values were computed for the all-vs-all superimposition with Fit3D [133]. Additionally, the RMSD differences between M1 and M2 for both motifs are given in respect to their backbone ($\Delta_{backbone}$) and side chain ($\Delta_{side\ chain}$) atoms. Due to the varying amino acid composition, the side chain RMSD values for the Backbone Brackets motif were computed by considering the last heavy side chain atom.

	binding mode	observations	$RMSD_{backbone}$ [Å]	$\Delta_{backbone}$ [Å]	$RMSD_{side\ chain}$ [Å]	$\Delta_{side\ chain}$ [Å]
Backbone Brackets	M1	28	0.32	0.02	1.42	0.00
	M2	59	0.34		1.42	
Arginine Tweezers	M1	39	0.24	0.04	0.72	0.66
	M2	47	0.28		1.38	

Backbone Brackets observations, with a mean of 17.92 ± 0.86 Å for M1 and 18.41 ± 0.82 Å for M2, respectively. However, some observations (structures PDB:5v0i chain A, PDB:1jzq chain A, PDB:3tzl chain A, and PDB:3ts1 chain A) exhibit higher alpha carbon distances of 20.54 Å, 19.74 Å, 19.10 Å, and 18.79 Å, respectively. In contrast, one occurrence of the Backbone Brackets motif in structure PDB:4aq7 chain A has a remarkably low alpha carbon distance of 16.50 Å. Nevertheless, alpha carbon distances between bound and unbound state differ significantly (p-value<0.01, Figure 5.7). This indicates the substantial contribution of backbone interactions as well as the conformational change observed during ATP binding. The side chain variation is marginal for the Arginine Tweezers if ATP is bound. In contrast, the side chain angle of the apo form is highly variable with a mean of 91.82 ± 8.69° for M1 and 79.81 ± 21.67° for M2, respectively. The side chain angles between the bound and unbound state differ significantly (p-value<0.01, Figure 5.8), reinforcing the pivotal role of highly specific side chain interactions during ligand binding. This effect cannot be observed for the alpha carbon distances of the Arginine Tweezers, with a mean of 14.76 ± 0.66 Å for M1 and 14.93 ± 0.79 Å for M2, respectively.

Ligand-Based Alignment In order to relate the position of Backbone Brackets and Arginine Tweezers to the different attack modes during the amino acid activation step [88], a ligand-based alignment was computed. Class II aaRSs bind ATP in a unique bent conformation [171], while ATP binds in Class I aaRSs in its usual extended form. As done by DUTTA ET AL., the adenine substructure served as the basis for the alignment [100]. Figure 5.9 shows the ligand-based alignment including the structures of both motifs. It is evident that both structural motifs seem to be "mirror images" of each other. While the Backbone Brackets are oriented diametrically in respect to the ligand (Figure 5.9A), the Arginine Tweezers are located around the ATP part of the ligand. More specifically, they attack the ATP ligand from both sides of a reflection plane defined by the adenine substructure (Figure 5.9B). In respect to the general orientation of the ligand in the binding pockets, the two distinct conformations of ATP for Class I and Class II aaRSs, extended and bent, are obvious (Figure 5.9C). Based on the results of the alignment noncovalent interactions and the mode of attack of both structural motifs were investigated (Figure 5.9D). The Backbone Brackets interact with the ligand via hydrogen bonds exclusively. One bond is formed between the oxygen atom of α-phosphate and the backbone nitrogen atom of the N-terminal residue. The C-terminal residue interacts with the primary amino group at the 6' position of the ring structure via the backbone oxygen atom. Hence, donor and acceptor roles are fulfilled by different atoms of the amino acid backbone: nitrogen as donor in case of the N-terminal residue and oxygen in case of the C-terminal residue. The Arginine Tweezers interact with the ligand via a mix of hydrogen bonds, π-cation interactions, and salt bridges. While the N-terminal residue preferably binds to the α-phosphate with a salt bridge, the C-terminal residue mediates interactions through a hydrogen bond to the hydroxy group at the 3' ribose position and a π-cation interaction to the pyrimidine ring of adenosine.

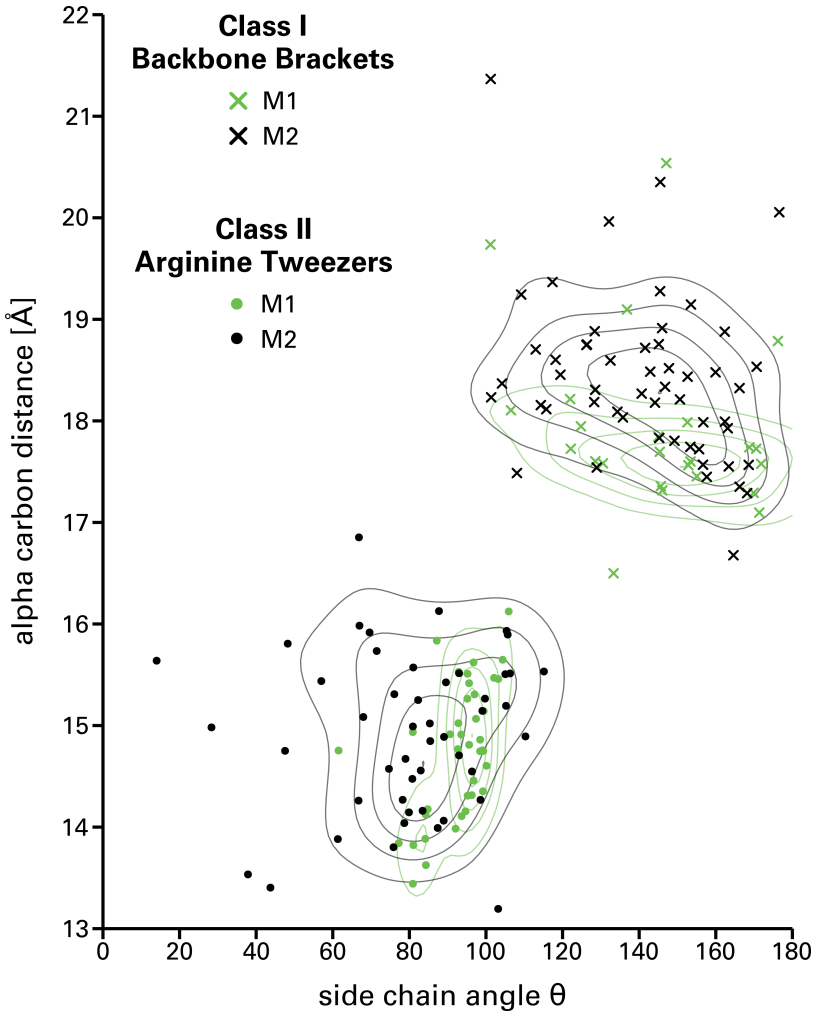


Figure 5.6.: The alpha carbon distance is plotted against the side chain angle θ. Binding modes refer to states containing an ATP ligand (M1) or not (M2). Backbone Brackets in M1 allow for minor variance with respect to their alpha carbon distance, constrained by the position of the bound ligand. In contrast, Arginine Tweezers in M1 adapt an orthogonal orientation in order to fixate the ligand.

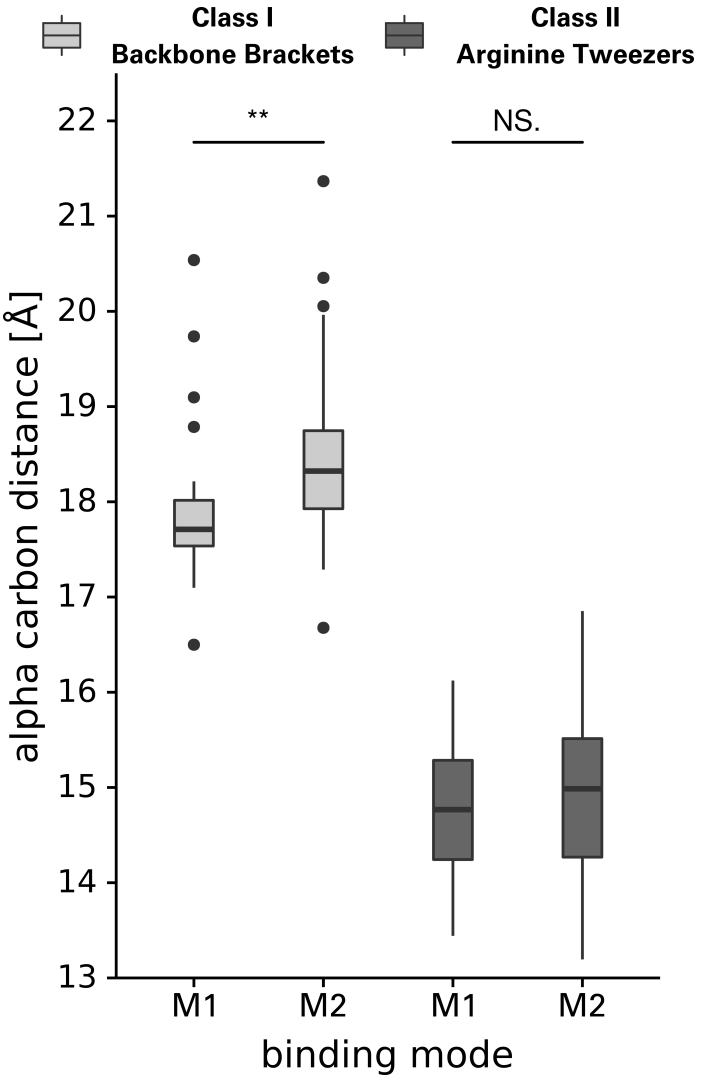


Figure 5.7.: Distributions of alpha carbon distances for Class I Backbone Brackets motif and Class II Arginine Tweezers motif in ATP bound (M1) and unbound state (M2). The alpha carbon distance of the Backbone Brackets differs significantly between the two states (Mann-Whitney U p-value<0.01)

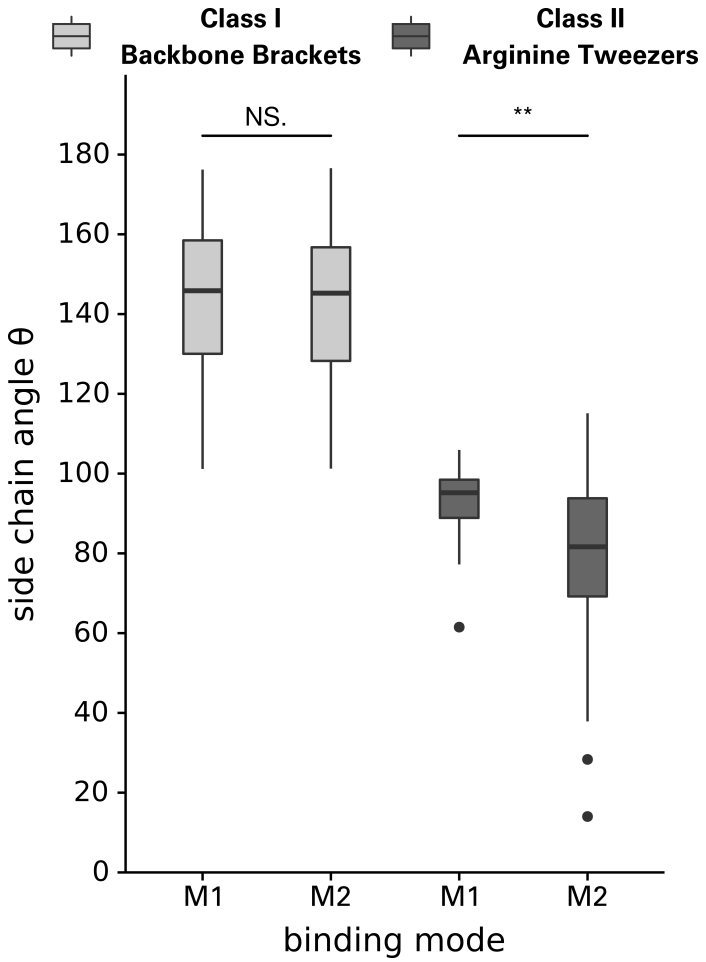


Figure 5.8.: Distributions of side chain angle θ for Class I Backbone Brackets motif and Class II Arginine Tweezers motif in ATP bound (M1) and unbound state (M2). The side chain angle of the Arginine Tweezers differs significantly between the two states (Mann-Whitney U p-value<0.01)

Interestingly, the N-terminal residues of both motifs are interacting with the α-phosphate, while the C-terminal residues form interactions with adenine and ribose. The N-terminal residues seem to be symmetric to the C-terminal residues in respect to a reflection plane defined by the ligand.

5.4.2. TEMPLATE-BASED DETECTION IN THE PROTEIN DATA BANK

In order to address the question, whether Backbone Brackets and Arginine Tweezers are generalizable patterns for ATP binding and if they can be observed in other ATP-binding proteins, a template-based detection with Fit3D [133] was performed across the whole PDB. This detection required the use of two core features of Fit3D: customizable atom representation and the definition of PSEs as described beforehand. The Backbone Brackets were represented by their backbone atoms, because these are the most constant part of this motif as shown in Section 5.4.1, while the Arginine Tweezers were represented by all atoms excluding hydrogen. The search database was created as described in Section 5.6 and all aaRS structures part of the dataset given in [163] were excluded. The top ten obtained matches of structural motifs similar to Backbone Brackets and Arginine Tweezers are listed in Tables 5.2 and 5.3, respectively.

Backbone Brackets In total, 607 matches of the Backbone Brackets motif were found in the PDB with a RMSD below 0.5 Å. Most of these matches can be attributed to unspecific matches. However, some of them show a striking similarity to the Backbone Brackets motif. A highly similar geometry with a RMSD of 0.1748 Å was found in an isocitrate dehydrogenase-2 enzyme (PDB:5kvu) of *Mycobacterium tuberculosis*. This class of enzyme catalyzes the oxidative decarboxylation of isocitrate to 2-oxoglutarate [172]. The reaction is dependent on $NADP^+$ as electron carrier, which contains an adenine-dinucleotide phosphate substructure; a structure similar to that of ATP. Closer investigation of the match in this enzyme reveals a highly similar binding pattern to that of Class I aaRSs (Figure 5.10A). Aspartic acid at position 605 binds the ligand using a hydrogen bond between the backbone oxygen atom and the primary amino group at the 6' position of the ring structure, whereas isoleucine at position 350 does not form any contacts to the α-phosphate but is in proximity to the β-phosphate of the ligand. Both, the C-terminal and N-terminal residue, feature characteristics similar to the Backbone Brackets motif and show a high geometric similarity. The backbone atoms of the N-terminal residue are oriented towards the ligand, potentially allowing the formation of hydrogen bonds. The assessment of other top-ranked matches remains difficult. The structures PDB:1f6d, PDB:1svv, PDB:5cuo, PDB:1ua4, PDB:5w7b, and PDB:2imz do not contain a ligand that is similar to, or contains the substructure of, ATP. Hence, no statement about generalizable similarity of the ligand recognition mechanism in these proteins to the Backbone Brackets motif can be made. However, the matches in PDB:3gag and PDB:4l2i do form interactions with the ligand. In the case of PDB:3gag, a nitroreductase-like protein, the residues 193 and 155, do barely interact with the flavin mononucleotide ligand. Only proline 155 forms hydrophobic interactions with the ligand. For PDB:4l2i, a flavoprotein, residues 267 and 306 feature a recognition mechanism similar to the Backbone Brackets. The N-terminal valine forms a hydrogen bond between its backbone oxygen atom and the proximal nitrogen atom of the flavin ring structure. Another hydrogen bond is formed between the nitrogen atom of the C-terminal aspartic acid and the hydroxy group at the 3' ribose position. This topology shows strong similarity to that of the Backbone Brackets motif. In the case of PDB:1w2l, a cytochrome C domain, interactions with the ligand are formed, but not via the backbones of the matched residues aspartic acid 53 and tyrosine 72.

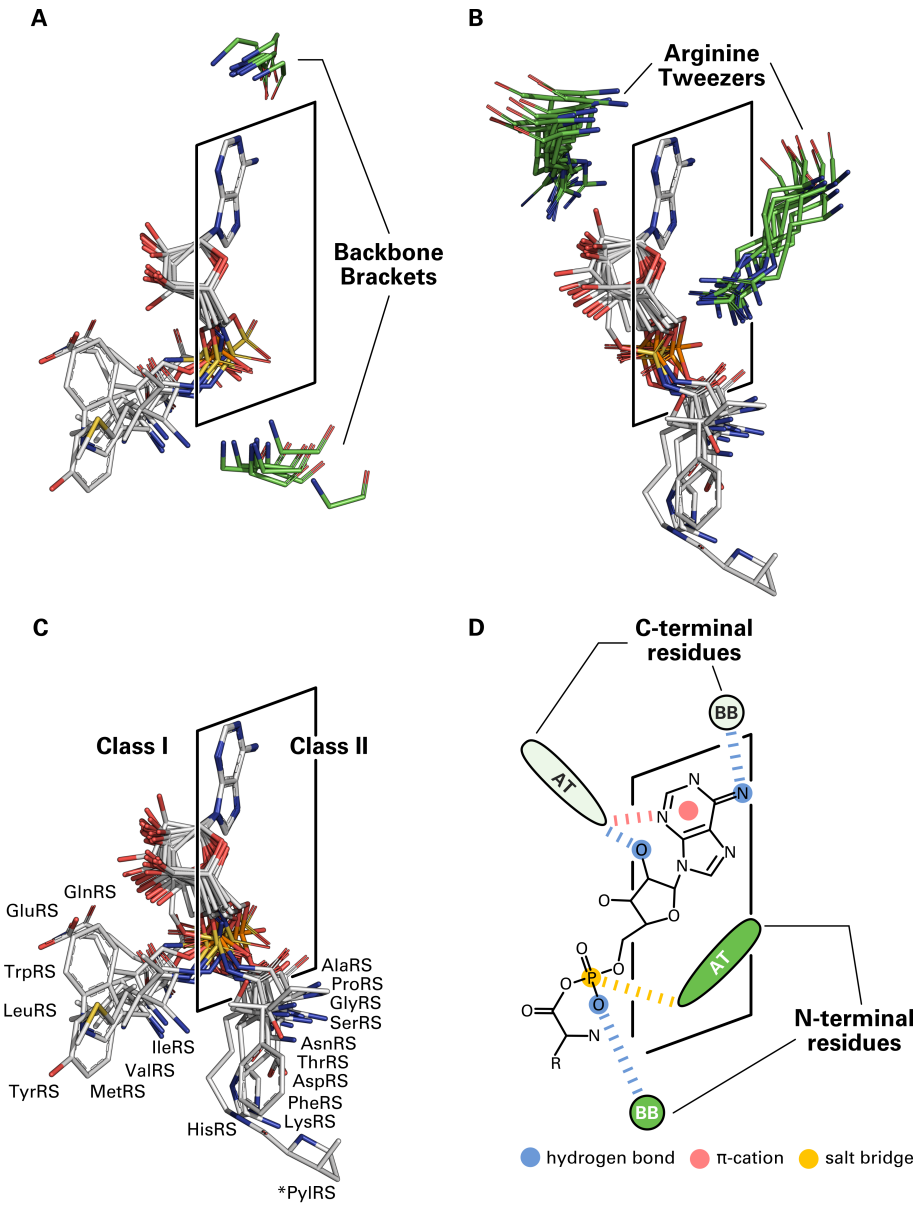


Figure 5.9.: Ligand-based alignment of the Backbone Brackets and Arginine Tweezers in respect to the adenine substructure of the ligand. Figure inspired by [100]. (**A**) Class I Backbone Brackets orientation in respect to the ligand after the ligand-based alignment. Both residues are oriented diametrically in respect to the ligand. Side chains of the residues are not shown. (**B**) Class II Arginine Tweezers orientation in respect to the ligand after the ligand-based alignment. Both residues are located around the ATP part of the ligand. Moreover, they are attacking the ligand from both sides of the reflection plane defined by the adenine substructure (black frame). (**C**) The two distinct conformations of the ligand, ATP in extended and bent form [100], for Class I and Class II aaRSs is evident. (**D**) Schematic depiction of the orientation of both motifs in respect to the ligand and the interactions they mediate (determined with PLIP [167]). While the Backbone Brackets (BB) form hydrogen bonds with the ligand, the Arginine Tweezers (AT) utilize hydrogen bonding, a π-cation interaction, and a salt bridge. The N-terminal residues of both motifs are directly interacting with the α-phosphate of ATP, while the C-terminal residues interact with adenine and ribose.

Table 5.2.: The top ten matches of the Backbone Brackets motif across the PDB computed with Fit3D [133]. Matches were filtered for redundancy using a sequence similarity cutoff of 95%. The statistical significance of the matches was assessed [162]. If available, mappings to UniProt [175] and Pfam [132] database, and the EC number [41] are given. Mappings were determined using the SIFTS project [176].

match	RMSD [Å]	significance	structure	UniProt	Pfam	EC
5kvu_A-605_A-350	0.1748	***	Isocitrate Dehydrogenase-2	O53611	PF03971	n/a
1f6d_A-352_A-209	0.2208	***	UDP-N-Acetylglucosamine 2-Epimerase	P27828	PF02350	5.1.3.14
1svv_A-221_A-40	0.2602	***	Threonine Aldolase	E9AC39	PF01212	n/a
3gag_A-193_A-155	0.2711	***	Nitroreductase-Like Protein	Q8DVW4	PF00881	n/a
5cuo_A-157_A-97	0.2782	***	Phosphotransacylase	Q21A54	PF06130	n/a
1ua4_A-342_A-30	0.2878	***	ADP-Dependent Glucokinase	Q9V2Z6	PF04587	2.7.1.147
4l2i_A-306_A-267	0.2884	***	Flavoprotein	D2RIQ3	PF01012	n/a
1w2l_A-53_A-72	0.2929	***	Oxidoreductase Cytochrome C Domain	Q9F3S9	PF00034	n/a
5w7b_C-356_C-232	0.2968	***	Acyloxyacyl Hydrolase	O18823	PF00657	3.1.1.77
2imz_A-70_A-425	0.2972	***	Mtu ReCa Intein Splicing Domain	P9WHJ3	n/a	n/a

*** *p*-value<0.001

Table 5.3.: The top ten matches of the Arginine Tweezers motif across the PDB computed with Fit3D [133]. Matches were filtered for redundancy using a sequence similarity cutoff of 95%. The statistical significance of the matches was assessed [162]. If available, mappings to UniProt [175] and Pfam [132] database, and the EC number [41] are given. Mappings were determined using the SIFTS project [176].

match	RMSD [Å]	significance	structure	UniProt	Pfam	EC
3g1z_A-100_A-303	0.2552	***	Potential tRNA Synthetase	Q9ZJ12	PF00152	n/a
6bni_A-295_A-523	0.3658	**	Lysyl-tRNA Synthetase	Q5CR27	PF01336	n/a
6chd_A-323_A-553	0.3755	**	Human Lysyl-tRNA Synthetase	Q15046	PF01336	6.1.1.6
6aqg_A-255_A-474	0.4588	*	Lysyl-tRNA Synthetase	A0PV47	PF01336	n/a
6aqh_A-258_A-478	0.5341	*	Lysyl-tRNA Synthetase	G7CF12	PF01336	n/a
4h2w_B-159_B-286	0.5645	*	Glycine Carrier Protein Ligase	Q89VT8	n/a	6.2.1.n2
12as_A-100_A-299	0.7089	N.S.	Asparagynyl-tRNA Synthetase	P00963	PF03590	6.3.1.1
6blj_A-306_A-442	0.7091	N.S	Seryl-tRNA Synthetase	n/a	n/a	n/a
3a5y_A-100_A-303	0.7446	N.S	GenX	A:P0A8N7	PF00152	n/a
4hvc_A-1152_A-1278	0.8148	N.S.	Prolyl-tRNA Synthetase	P07814	PF09180	6.1.1.17

* *p*-value<0.1, ** *p*-value<0.01, *** *p*-value<0.001

Arginine Tweezers For the Arginine Tweezers motif, 158 matches were found in the PDB with a RMSD below 1.5 Å. Eight of the ten top-ranked matches (PDB:3g1z, PDB:6bni, PDB:6chd, PDB:6aqg, PDB:6aqh, PDB:12as, PDB:6blj, and PDB:4hvc) were found in structures of aaRSs. These structures were not contained in the original dataset and thus not filtered before the template-based screening. One match was found in an amino acid:[carrier protein] ligase (aa:CP) of *Agrobacterium tumefaciens* (PDB:4h2w) and in a protein that aminoacylates the translation elongation factor P (EF-P) in *Escherichia coli* with lysine (PDB:3a5y). Both of these proteins show a high similarity to Class II aaRSs. The EF-P, which is aminoacylated by the protein PDB:3a5y, mimics the shape of a tRNA molecule and its aminoacylation site resembles the tRNA 3' end [173]. The structure contains an ATP ligand, identical to that of native aaRSs. Both matched residues, 100 and 303 in chain A, interact with the ATP ligand in a way identical to the Arginine Tweezers motif via hydrogen bonds, salt bridges, and π-cation interactions. Aa:CPs are another remarkable class of enzymes, responsible for the transfer of amino acids to the phosphopantetheine group of small carrier proteins [174]. In the case of the match in the aa:CP in structure PDB:4h2w, the residues 159 and 186 interact with the ligand via salt bridges, and π-cation interactions identically to that of the Arginine Tweezers (see Figure 5.10B).

5.4.3. TEMPLATE-FREE DETECTION

In addition to a template-based detection in order to identify structural motifs similar to the Backbone Brackets and Arginine Tweezers, the capability of Fit3D for the template-free detection of substructure similarity was applied to all Class I and Class II aaRS struc-

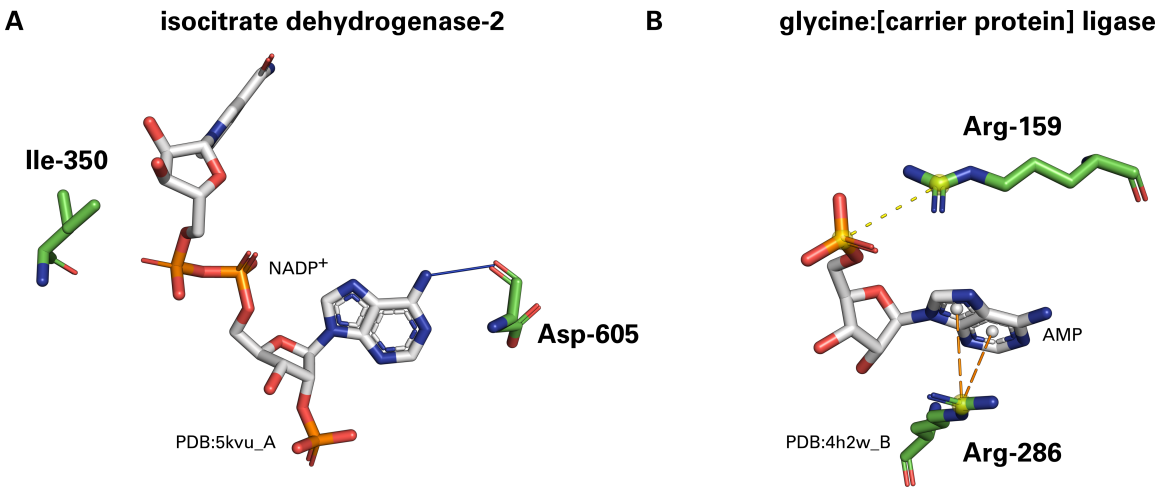


Figure 5.10.: Structural motifs similar to Backbone Brackets and Arginine Tweezers found in the Protein Data Bank with Fit3D [133]. (**A**) The structure of a binding motif highly similar to the Backbone Brackets. The match was found in an isocitrate dehydrogenase-2 enzyme (PDB:5kvz chain A) of *Mycobacterium tuberculosis*. The C-terminal motif residue aspartic acid 605 forms a hydrogen bond (determined with PLIP [167], solid blue line) with the $NADP^+$ ligand. (**B**) The structure of a binding motif highly similar to the Arginine Tweezers found in glycine:[carrier protein] ligase structure PDB:4h2w chain B. The interactions formed with the AMP ligand are similar to those in the Arginine Tweezers motif. While the N-terminal residue utilizes salt bridge interactions with the α-phosphate to fixate the ligand (yellow dashed lines), the C-terminal residue forms π-cation interactions with the ring systems of the adenine substructure (orange dashed lines).

tures, respectively. Without providing any *a priori* knowledge, the aim was to identify other hotspots in the structures of aaRSs that feature an outstanding structural conservation at atomic level. For this purpose the developed Fit3D algorithm was used, presented in detail in Section 6.2. The chosen parameters and methodologies are elucidated in Section 5.6. The detection process included inter-molecular interaction data as determined with PLIP [167]. A classification of amino acids according to their chemical groups[3] was used as proposed by GUTTERIDGE AND THORNTON to cover isofunctional mutations [42]. For visualization purposes, a representative structure was selected that is used to highlight conserved substructures. The relative coverage of determined structural motifs for each position in the structure is depicted by color intensity and loop thickness (see Section 6.2.2 on how the coverage is calculated).

Class I Structures All 81 non-redundant Class I structures of the dataset presented in [163] were used for the template-free detection of conserved substructures. Figure 5.11 shows the results of the template-free detection algorithm. An ArgRS structure (PDB:1f7u) from *Saccharomyces cerevisiae* is used as the reference for visual representation. Red and bulky areas in the structure depiction correspond to residues of high structural conservation, i.e. geometrically conserved structural motifs. Histidine 159 and histidine 162 are salient and exhibit a high coverage of 0.19 and 1.00, respectively (Figure 5.11A). The corresponding itemset hyb-pic-imi-oth (p-value<0.001, Table A.1), which represents the structural motif, features two populations of geometrically highly conserved histidine residues, a π-cation interaction, and a hydrogen bond (Figure 5.11B). Both residues are located in the catalytic core domain of Class I aaRSs and constitute the so-called HIGH motif [11], which is known to stabilize the transition state during the aminoacylation reaction and is thus highly relevant for catalysis. Other structurally conserved residues with known bio-

[3] **imidazole (imi)**: histidine; **guanidinium (gua)**: arginine; **amine (amn)**: lysine; **carboxylate (car)**: aspartic acid, glutamic acid; **amide (amd)**: asparagine, glutamine; **hydroxyl (hyd)**: serine, threonine, tyrosine; **thiol (thi)**: cysteine; **others (oth)**

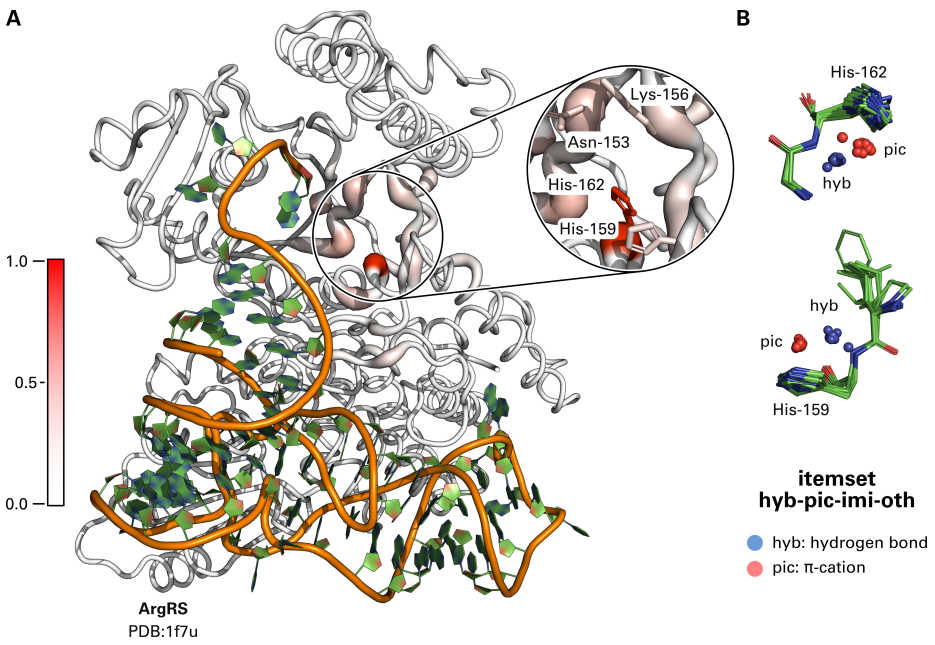


Figure 5.11.: Results of the template-free structural motif detection in Class I aaRSs. (**A**) An ArgRS structure (PDB:1f7u) from *Saccharomyces cerevisiae* in complex with tRNA. The color intensity and thickness of the loop representation of the structure corresponds to the coverage of conserved structural motifs at a certain residue position. Histidine residues at positions 159 and 162 exhibit a high structural conservation across all Class I structures. (**B**) The corresponding itemset hyb-pic-imi-oth that contains both histidine residues. Small spheres represent the midpoint of an interaction determined with PLIP [167].

Table 5.4.: The structural conservation of individual residues in aaRS Class I structure PDB:1f7u as determined with the template-free structural motif detection algorithm [177]. Only residues with a biological role are shown. The residues are sorted according to a descending coverage score.

residue type	position	coverage	role	reference
His	162	1.00	C-terminal HIGH motif residue	[11]
Asn	153	0.23	N-terminal Backbone Brackets residue	[163]
Lys	156	0.17	phosphate binding site	[163]
His	159	0.14	N-terminal HIGH motif residue	[11]

logical role are asparagine at position 153 with a coverage of 0.23 and lysine 156 with a coverage of 0.17, respectively. Asparagine 153 corresponds to the N-terminal residue of the Backbone Brackets motif [163]. Lysine 156 participates in binding of the phosphate part of the ATP ligand in ArgRS by establishing hydrogen bonds and salt bridges [163]. All functionally relevant residues of structure PDB:1f7u backed by geometrically conserved structural motifs are listed in Table 5.4.

Class II Structures The template-free detection in all 76 non-redundant Class II aaRSs structures did not yield comparable results to that of Class I. Residue coverage of the reference structure PDB:1c0a chain A, an AspRS from *Escherichia coli*, is given at some positions in the catalytic core domain but the corresponding itemsets do not show a high geometric conservation. The best scoring itemset hyb-pis-imi-oth (see Table A.2) has a consensus score of 0.3565 and 54 evenly populated structural clusters compared to a consensus score of 0.2381 and 29 clusters (of which few are highly populated) for the best scoring itemset hyb-pic-imi-oth in Class I aaRSs.

5.5. DISCUSSION

The reflexive system of building blocks and building machinery implemented in aaRSs is an intriguing aspect of the early development of living systems. There is evidence that proteins arose from an ancient set of peptides [178] and that these peptides were co-factors of the early genetic information processing by RNA. The Backbone Brackets and the Arginine Tweezers are an outstanding example how nature realized the binding of the same ligand species with completely different mechanisms. A mechanism similar to the Backbone Brackets and Arginine Tweezers motif was observed in carbohydrate kinases, where an isofunctional ligand binding mechanism is used among a heterogeneous set of proteins [179]. Another example of highly variable implementations of ligand binding are non-ribosomal peptide synthetases as another enzyme family that is required to recognize all 20 amino acids [180].

Relevance of Structure Studies Sequence-based analyses were among the first tools to investigate the transfer of genetic information. DNA and protein sequences comprise the developmental history of organisms, their specialization, and diversification [71]. However, following the "functionalist" principle in biology, sequence is less conserved than structure, which is, in turn, less conserved than function [4]. Therefore, structural features and molecular contacts have been recognized as key aspects in grasping protein function [42, 168] and evolution. Only if the necessary function can be maintained by compatible interaction architectures, the global role of the protein in the complex cellular system is ensured [5]. This is also eminent in aaRS precursor structures that were described to be molten globules but as long as the function of the protein is ensured, it is able to survive during evolution [19]. If evolution tries to conserve structure over function, the evolutionary progress might have been considerably slower and thresholds for the development of new functions would have been higher [4]. Each amino acid of a protein fulfills a certain role and can often be replaced by amino acids with compatible attributes [42]. By considering each amino acid in the context of its sequence, its structural surroundings, and finally its biological function, one can determine possible exchanges and the evolutionary pressure driving these changes [4, 181]. Up to this point, pure sequence analysis methods – ignoring structural motifs and ligand interaction data – missed the functional relevance of the Backbone Brackets entirely. By applying the structural motif detection algorithms presented in this thesis, detailed insights into the characteristics of the two structural motifs were gained.

5.5.1. STRUCTURAL CHARACTERIZATION

The results of the structural characterization allowed to dissect the molecular ATP recognition mechanism at atomic level. The application of Fit3D enabled the quantification of structural changes in the binding site of aaRSs upon ligand binding.

Geometric Analysis The geometric characterization of the two structural motifs (see Figure 5.6) highlighted some observations of the Backbone Brackets, which exhibit a substantial increase or decrease of the residue alpha carbon distance. For instance, chain A of an leucyl-tRNA synthetase (LeuRS) of *Escherichia coli* (PDB:4aq7) is complexed with tRNA and the Backbone Brackets alpha carbon distance is about one Ångstroem below the average. Manual investigation of this structure showed that there is no obvious conformational difference to other structures. Likewise, the annotated interactions were checked for consistency using PLIP and showed usual interactions with the adenine and the sulfamate group (the phosphate analogue) of the ligand. For the Backbone Brackets with higher

alpha carbon extent, structures of isoleucyl-tRNA synthetase (IleRS), TrpRS, and tyrosyl-tRNA synthetase (TyrRS), interaction analysis revealed that residue 274 interacts with the amino acid side chain, as all of these structures contain a single aminoacyl ligand (PDB:3tzl chain A, PDB:3ts1 chain A, PDB:1jzq chain A) or two separate ligands (amino acid and AMP, PDB:5v0i chain A). This suggests that the structures resemble a partially changed conformation prior to tRNA ligation and a possible role of the Backbone Brackets motif in amino acid recognition. Likewise, these effects can arise from low quality electron density maps in the structure regions of interest. However, these hypotheses have to be addressed and validated in future work.

Backbone Brackets Indicate a Conformational Change The geometric analysis of the Backbone Brackets motif (Section 5.4.1) showed a high variance of side chain angles for both binding modes. The distinction between these modes is significantly manifested in a change of the alpha carbon distance, which supports that the conformational change during ligand binding previously observed in ArgRSs [182], TyrRSs [73–76, 183], and TrpRSs [77, 79–81] is a general mechanism in Class I aaRSs. Furthermore, the C-terminal residue of the Backbone Brackets is located close to the KMSKS sequence motif [163]. Thus, the structural rearrangement in the KMSKS motif upon ATP binding might indirectly affect the geometric orientation of the C-terminal residue of the Backbone Brackets – especially regarding the position of its alpha carbon atom.

Arginine Tweezers Form Highly Specific Side Chain Interactions In contrast to the Backbone Brackets, the Arginine Tweezers are highly restrained in side chain orientation if a ligand is bound, which shows that correct orientation is key for proper ATP recognition. In principle, π-cation interactions between aromatic ligands and binding site residues can be formed with any of the positively charged residues lysine, arginine, or histidine. However, π-cation interactions between aromatic ligands and arginine were shown to be robust in polar surroundings [184]. Thus, it might have been beneficial for aaRSs Urzymes, which formed molten globules [19, 22] without ordered tertiary structure, to preserve interactions in the ancient polar environment by utilizing arginine. The Arginine Tweezers geometry is less limited, which is reflected in a higher variability of side chain orientations if no ligand is bound. The flexibility of arginine side chains observed in aaRSs is in conjunction with previous studies [185]. A previously observed conformational change in the active site of SerRS [186, 187] upon ligand binding could not be confirmed to be a general mechanism in Class II. For the structures in the non-redundant dataset [163] the alpha carbon distance does not differ significantly between bound and unbound state. However, the distinction between the two binding modes can be made by taking the geometry of the two motifs into account: alpha carbon distances for Backbone Brackets and side chain angles for Arginine Tweezers.

Backbone Brackets and Arginine Tweezers are Mirror Images of Each Other The results of the ligand-based alignment (Section 5.4.1, Figure 5.9) allowed to relate the orientation of the Backbone Brackets and Arginine Tweezers to each other. In conjunction with the results of DUTTA ET AL., the analysis shows the different modes of attack for Class I and Class II aaRSs [100], respectively. Remarkably, the C-terminal residues of both structural motifs interact with the adenine and ribose substructures exclusively. Both N-terminal residues are located in the Urzyme region [163] proposed by the Rodin-Ohno hypothesis, suggesting the evolutionary trajectory of ATP ligand recognition started with specific interactions to the adenine and ribose part of the ligand. Consequently, interactions with the α-phosphate are presumably younger and occurred later during aaRS evolution. The

C-terminal Backbone Brackets residue is located in the Urzyme region as well [163], suggesting the salt bridge interactions between the C-terminal Arginine Tweezers residue and the ligand to be the youngest. In general, both motifs seem to be mirror images of each other regarding a reflection plane defined by the ATP ligand. This is underlined by the fact that both motifs feature residues for ligand recognition at both moieties of the ligand: the adenine as well ribose substructures and the α-phosphate, respectively.

5.5.2. TEMPLATE-BASED DETECTION IN THE PROTEIN DATA BANK

Backbone Hydrogen Bonds are a General Mechanism The Backbone Brackets are remarkable structural motif, since backbone interactions are often neglected in structural studies. Nevertheless, backbone hydrogen bonds make up at least one quarter of overall ligand hydrogen bonding [188]. In these cases, side chain properties may only play a minor role, e.g. for steric effects, and allow for larger flexibility in implementation of a binding pattern as long as the correct backbone orientation is ensured. The Backbone Brackets motif is a prime example for conservation of function over structure or sequence [4]. When ligands can still be bound specifically by backbone interactions, these binding sites become significantly more resilient to mutations. There are other examples of protein-ligand complexes where backbone hydrogen bonds are a major part of the binding mechanism, e.g. in binding of the cofactor NAD to a CysG protein from *Salmonella enterica* (PDB:1pjs) as determined with PLIP [167]. In conclusion, the Backbone Brackets exhibit conservation at a functional level rather than at the sequence level, which renders sequence-based motif analysis infeasible. The feature of Fit3D [133] to define allowed isofunctional mutations (PSEs) at each position of the template motif individually, allows to tackle this problem at a structural level.

The results of the template-based detection in the PDB unveiled structural motifs with a remarkable similar geometry to the Backbone Brackets. However, only the top-scoring match in structure PDB:5kvu shows a functional similarity to the Backbone Brackets motif. It can be assumed that the binding of $NADP^+$ in this structure is accomplished by a mechanism similar to that of the Backbone Brackets in Class I aaRSs. In structure PDB:4l2i, where a similar geometry binds to the FAD ligand via backbone hydrogen bonds, valine 267 has been shown to be involved in ligand binding [189]. Along with the results obtained for Class I aaRSs, this emphasizes the important role of backbone hydrogen bonds. However, especially due to the high structural variation of the Backbone Brackets in terms of alpha carbon distances, the results of the detection include false positive matches, e.g. structures PDB:156d or PDB:1svv where similarity is likely due to chance. This also relates to the effect of scoring the similarity with the RMSD that is highly dependent on the count of atoms used for alignment [162]. Due to their variability in protein sequence, the Backbone Brackets motif can only be represented by its eight heavy atoms of the amino acid backbone which leads to a decrease in sensitivity and specificity of the detection method.

Arginine Tweezers are Unique for Aminoacyl-tRNA Synthetases and Paralogs The conserved Arginine Tweezers motif resembles a common interaction pattern for phosphate recognition [42], which usually features positively charged amino acids [190]. However, the conformational space of ATP ligands was shown to be large throughout diverse superfamilies [171] and hence the geometry of binding sites involved in ATP recognition is manifold. The uniqueness of aaRSs compared to other ATP-binding proteins was demonstrated for AspRS, where the ligand binds in a compact form with a bent phosphate tail instead of the usually found extended form [171]. This conformation of ATP is energetically unfavorable but allows easy access of the α-phosphate for tRNA binding [191]. In general, the nucleophilic attack to the α-phosphate of ATP is oppositely directed in Class I and Class II aaRSs

which possibly evolved at pre-biotic time [100]. Quantum mechanical calculations have shown that a lesser propensity for the nucleophilic attack of Class II amino acids is compensated by the bent state of ATP, related binding site residues, and magnesium ions [100]. This specialized mechanism in Class II aaRSs suggests that the Arginine Tweezers motif possesses a unique geometry and is not a generalizable pattern for ATP binding, such as the frequently occurring P-loop domain [190]. The results of the template-based detection for similar structural motifs throughout the PDB (Section 6.1) support this assumption. Seven of the top ten matches reported by Fit3D occurred in functional Class II aaRS structures, e.g. a human LysRS (PDB:6chd). Only two matches were reported in other enzymes that are, however, aaRS Class II paralogs. The top-scoring match was reported in a structure with putative aaRS function [192]. In the case of the seven functional aaRS structures, these were not part of the original dataset [163]. For example, the human LysRS structure PDB:6chd was released in November 2017; it too recent to be included in the original dataset. The same holds true for the structures PDB:6bni, PDB:6aqg, PDB:6aqh, and PDB:6blj. Structure PDB:4hvc, a ProRS, was released in 2012 and thus was obviously missed in original dataset, probably because its EC number EC:6.1.1.17 was not considered for initial dataset creation (see [163] Supporting Information). This demonstrates how high-precision template-based structural motif detection allows identifying other structures with the same function beyond database annotation.

The top-ranked structure PDB:3g1z was shown to catalyze a post-translational modification of EF-P [192], which is a mechanism similar to that shown for PDB:3a5y [173]. Despite the lack of the tRNA aminoacylation capability of these enzymes [192], they are still able to catalyze this reaction for EF-P. Because both of the structures contain the Arginine Tweezers motif, this again suggests its necessity for catalytic activity. A remarkable occurrence of a structural motif similar to the Arginine Tweezers occurs in aa:CP enzymes. These enzymes show a high similarity to Class II aaRSs and are required to recognize amino acids specifically but are unable to aminoacylate tRNA [174]. Because of their inability to charge tRNA but small carrier proteins, aa:CP might allow an insight into the evolutionary period before tRNA recognition of aaRSs evolved [174]. This again emphasizes the coupling of the Arginine Tweezers and the evolution of ATP recognition in aaRSs.

In general, the Arginine Tweezers motif seems to be a unique structural motif, which exclusively occurs in aaRSs or close paralogs. This is in coherence with the energetically unfavorable bent state of the ATP ligand [171], which might required specialized recognition mechanisms to evolve to keep the efficiency of Class II aaRSs on par with that of Class I [100].

5.5.3. TEMPLATE-FREE DETECTION

The template-free detection of geometrically conserved structural motif in aaRSs resulted in a high agreement between geometric conservation and functional importance for Class I but not for Class II aaRSs. Template-free detection in Class II aaRSs did not pinpoint any generalizable structural motifs with a strong geometric conservation. In general, the frequent domain inserts [19, 22] occurring in Class II aaRSs may influence the detection algorithm negatively. Additionally, the weak conservation [59] of Class II motifs and the variability in their relative arrangement [64] are not beneficial for the conservation of small substructures. The Arginine Tweezers were shown to be variable in their side chain orientation if no ligand is bound (see Section 5.4.1), which can disturb the high-precision detection that considers also the orientation of side chains. Further analysis of Class II aaRSs should probably be applied on the structurally conserved Urzyme region and ATP bound structures exclusively.

However, in Class I aaRSs a structural motif was detected that includes the N- and

C-terminal residues of the HIGH motif. This motif is of great importance for catalysis in Class I aaRSs. For TyrRS, mutations of any histidine of the HIGH motif [72] have been shown to decrease activity, since both residues contribute to the stabilization of the transition state of the reaction [193,194]. This coincides with the strong geometric conservation observed for the these residues. It seems that a π-cation interaction between both histidine residues is a determinant for the rigidity of the HIGH motif residues. This "parking position" of the HIGH motif might explain why only a few irregularly occurring interactions with the ligand have been observed in the pre- and post-aminoacylation state [163]. The application of a template-free structural motif detection algorithm allowed to pinpoint and generalize the strong geometric conservation of the HIGH motif for the first time. Other important positions which show substructure conservation include, for example, asparagine 153 and lysine 156 in structure PDB:1f7u. Because these residues were shown to participate in ATP binding [163], this emphasizes the suitability of template-free structural motif detection to discover functionally relevant structural motifs. Asparagine 153 constitutes the N-terminal residue of the Backbone Brackets motif.

5.6. MATERIALS AND METHODS

Structural Motif Alignments and Geometric Analysis All motif occurrences in M1 and M2 representative chains as defined in [163] were aligned in respect to their backbone atoms using the Fit3D algorithm [195]. Additionally, the alpha carbon distances and the angle between side chains were determined. The side chain angle θ between two residues was calculated by abstracting each side chain as a vector between alpha carbon and the most distant carbon side chain atom. Hence, the side chains are oriented in a parallel way if θ=0° or θ=180°. Side chain angles were not calculated if one or both residues of a Backbone Brackets observation were glycine.

Ligand-Based Alignment All non-redundant structures, representative for ATP binding (determined as described in [163] Supporting Information), were used for the ligand-based alignment. For each of these structures only the ligand and both residues of the Backbone Brackets and Arginine Tweezers motif, respectively, were considered. These reduced structures where then aligned in respect to the adenine substructure of ATP by using the Fit3D API version [163]. The correct pairing of atoms was determined with subgraph isomorphism detection [46], where the adenine substructure served as pattern graph and the complete ATP ligand of each structure as target graph.

Template-Based Detection in the Protein Data Bank The template-based detection of Backbone Brackets and Arginine Tweezers in the PDB was based on a snapshot of the PDB as of April 19, 2018. All ligand-containing structures were retrieved with the advanced search functionality[4] of the PDB. The constraints were set to structures that contain at least one ligand. Subsequently, a customizable table report of the search results was generated, which included the entity identifier of the structure (corresponds to a single protein chain). This resulted in 332,319 individual macromolecular chains, including other types of macromolecules such as DNA and RNA. For all of these chains, ligand interaction data were calculated with the PLIP command line tool version 1.4.0 [167] and default setting. Only residues which were annotated to be in contact with any ligand were kept, non-interacting residues and other macromolecules than proteins were removed from the structures. This resulted in binding site data for 217,845 proteins. The template-based detection was then performed on this dataset of structures as follows. All occurrences of Backbone Brackets

[4] rcsb.org/pdb/search/advSearch.do?search=new, available as of April 19, 2018

(Arginine Tweezers) of Class I (Class II) structures, representative for ATP binding [163], were clustered using affinity propagation [196] with an epoch limit of 1,000, λ=0.50, and the self-similarity of two data points set to the inverse RMSD after superimposition. For Backbone Brackets only backbone atoms were considered, while all atoms excluding hydrogen were considered for the Arginine Tweezers. For all structural motifs reported to be exemplars according to the affinity propagation clustering, a template-based detection with Fit3D [133] was performed against the binding site data. For the Backbone Brackets PSEs against all other residues were defined. The RMSD cutoff was set to 0.50 Å for the Backbone Brackets and 1.50 Å for the Arginine Tweezers, respectively, which approximately corresponds to the average all-against-all RMSD values of these motifs (Table 5.1). After obtaining all matches, the results were filtered for redundancy using the PDB REST endpoint for sequence clusters[5] with a similarity cutoff of 95% sequence identity.

Template-Free Detection For Class I aaRS structures, the parameters of the template-free detection algorithm (see Section 6.2) were set as follows: maximal support of 0.80, maximal cohesion of 5.00 Å, maximal separation of 100.00 and optimal separation of 5.00, maximal consensus of 1.00 with λ=0.50. The Class I dataset contained 81 non-redundant structures, representative for sequence clusters with an identity of 95% and included PDB:1f7u for visualization of the coverage score. Residues were categorized according to their chemical groups [42]. Furthermore, structures were annotated with inter-molecular interaction data with each interaction represented by a pseudoatom at the midpoint of interacting atoms [197]. The interaction data was calculated with the PLIP command line tool version 1.4.0 [167] and the "`--intra`" command line flag enabled. The statistical significance of matches was calculated in respect to the consensus score (see Section 6.2). The parameters for 76 Class II structures, representative for sequence clusters with an identity of 95% and including PDB:1c0A for coverage visualization, were set identical to those for Class I.

[5] rcsb.org/pdb/rest/sequenceCluster, available as of April 19, 2018

6. FIT3D: STRUCTURAL MOTIF DETECTION ALGORITHMS

This chapter is based on the results of the articles "A novel algorithm for enhanced structural motif matching in proteins" published in *Journal of Computational Biology*, "Fit3D: a web application for highly accurate screening of spatial residue patterns in protein structure data" published in *Bioinformatics*, and "Unsupervised Discovery of Geometrically Common Structural Motifs and Long-Range Contacts in Protein 3D Structures" published in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. For a detailed list of author contributions please refer to page 17.

The following chapter presents Fit3D that provides two solutions for structural motif detection. First, a template-based approach [195] is discussed that is backed by a combinatorial algorithm. Second, a template-free method [177] is presented that exploits itemset mining to discover geometrically conserved structural motifs. Both methods are a major part of the contribution of this thesis and address the limitations of previous work that are described in Section 4.4. Beside the algorithmic concept, benchmarks and validations are provided.

6.1. TEMPLATE-BASED STRUCTURAL MOTIF DETECTION

One core feature of Fit3D is the accurate detection of matches of structural motifs based on a given template. This section relates the origin of this problem to computer science and describes an efficient algorithm to solve the problem. It can be applied to data representing macromolecular structures such as proteins but also to other types of structures, e.g. DNA or RNA.

6.1.1. PROBLEM ORIGIN

The detection of local similarities between a given template structural motif and substructures of other proteins is a pattern matching problem in the three-dimensional space. In the case of protein structures the data is usually labeled because each amino acid has an associated residue type such as alanine or glycine. Hence, the computational problem is to find the best agreement between a given set of labeled points, which represent the template motif, and sets of candidate points in the three-dimensional space. To find the best agreement two constraints have to be considered (Figure 6.1):

- the compatibility of the labels between template motif and the match candidate, and
- the RMSD of the points after optimal superimposition.

The problem of template-based structural motif detection is related to finding cliques in an undirected graph if the protein structure is represented as graph with residues as vertices and edges that represent contacts between residues with a defined distance cut-off. This strategy was used in the ProBiS algorithm [127, 198]. More general, the problem relates to subgraph isomorphism [46, 47]. Thus, it can be assumed that no exact solution exists that solves the problem of template-based structural motif detection in polynomial time. However, as shown later, filtering steps based on biological constraints are used, which allow the presented algorithm to run in applicable time.

In order to assess the geometric similarity of match candidates in respect to the template motif a transformation, consisting of a translation and rotation, has to be found that minimizes the RMSD. A popular algorithm that can be used for this purpose was described by KABSCH in 1978 [136]. It uses singular-value decomposition (SVD) to find a proper rotation that minimizes the RMSD between two point sets of identical size in the three-dimensional space. More recently, other approaches were presented that use the mathematical concept of quaternions to find the ideal superimposition [199, 200].

6.1.2. ALGORITHM

The following elucidations describe the algorithmic procedure of template-based structural motif detection with Fit3D. A corresponding pseudocode listing of the algorithm is given in Algorithm 1.

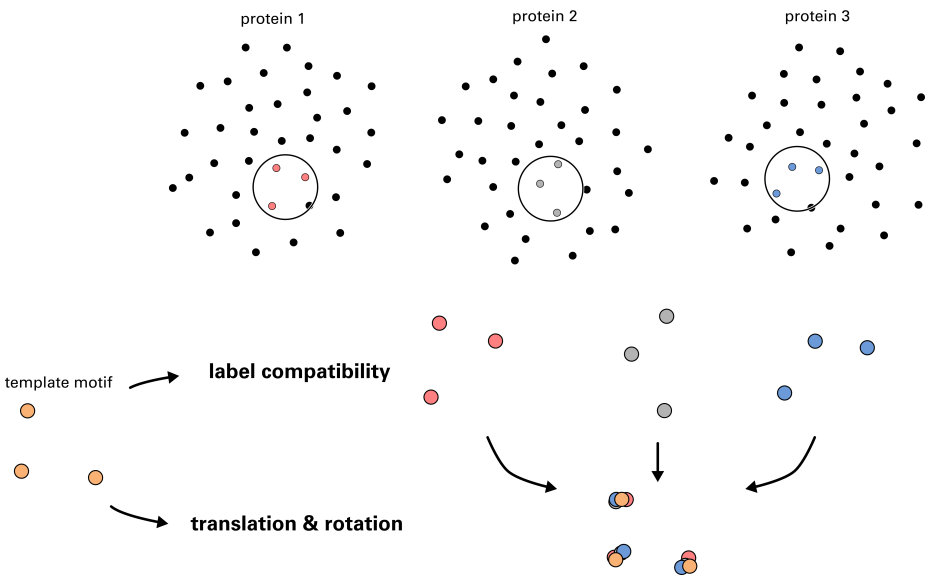


Figure 6.1.: Template-based structural motif detection can be seen as pattern matching in the three-dimensional space. Protein structures are represented as points. In order to detect matches in a given set of protein structures two constraints have to be evaluated for match candidates: the compatibility of labels (the types of residues) and the geometric similarity (e.g. RMSD) after translation and rotation defined by the optimal superimposition.

Assumptions The set $U = \{\text{Ala}, \text{Arg}, \text{Asn}, ..., \text{Val}\}$ corresponds to the three-letter code labels of all twenty canonical amino acid, hence $|U| = 20$ and $A = \{a_1, a_2, \ldots a_k\}$ is a set of amino acids. Let $f : A \rightarrow \mathcal{P}(U) \setminus \varnothing$ be a mapping function of amino acids to allowed labels. $\mathcal{P}(U)\setminus\varnothing$ corresponds to the power set of all labels excluding the empty set. In other words, the function f can describe both, the unambiguous residue – a residue is always compatible to itself – but also several other allowed residue types which correspond to PSEs. Further, let $g : A \rightarrow U$ be a mapping function solely representing the unambiguous and invariant assignment of amino acids to their labels. A template structural motif can be described as a set of amino acids and their corresponding atom coordinates. For simplification a computational representation of amino acids with a single atom (e.g. the alpha carbon atom) is assumed. Hence, the set $Q = \{q_1, q_2, \ldots, q_k\}$ represents a template structural motif consisting of k amino acids. Each amino acid $q = (g(q), f(q), v)$ is a triple with a label $g(q) \in U$, allowed PSEs $f(q)$, and coordinate information $v \in \mathbb{R}^3$. A target structure T consisting of n amino acids, in which a similar occurrence of the template structural motif has to be found, is represented similarly: $T = \{t_1, t_2, ..., t_n\}$ with $t = (g(t), v)$ being a tuple. A further input of the algorithm is the minimal required geometric similarity up to which matches should be reported. This corresponds to an upper bound ε for the RMSD dissimilarity measure after optimal superimposition of a match candidate and the template structural motif using the Kabsch algorithm [136].

Local Environments Fit3D is a combinatorial method that reduces the search space by considering so-called local environments, which are filtered according to several constraints. The local environment around an amino acid $t \in T$ of the target structure is denoted as E^t. Furthermore, let C be a set of match candidates for which the geometric similarity to the template motif Q should be evaluated. The maximal spatial extent r of the template motif, which is required for the extraction of local environments, is defined as the maximum of all pairwise distances between all alpha carbon atoms of the template motif: $r \leftarrow \max(\|q_i - q_j\| : q_i, q_j \in Q, i \neq j)$. Figure 6.2A exemplarily shows the determination of r and, for the case of the ES superfamily template, the mapping $f(q)\ \forall\ q \in Q$.

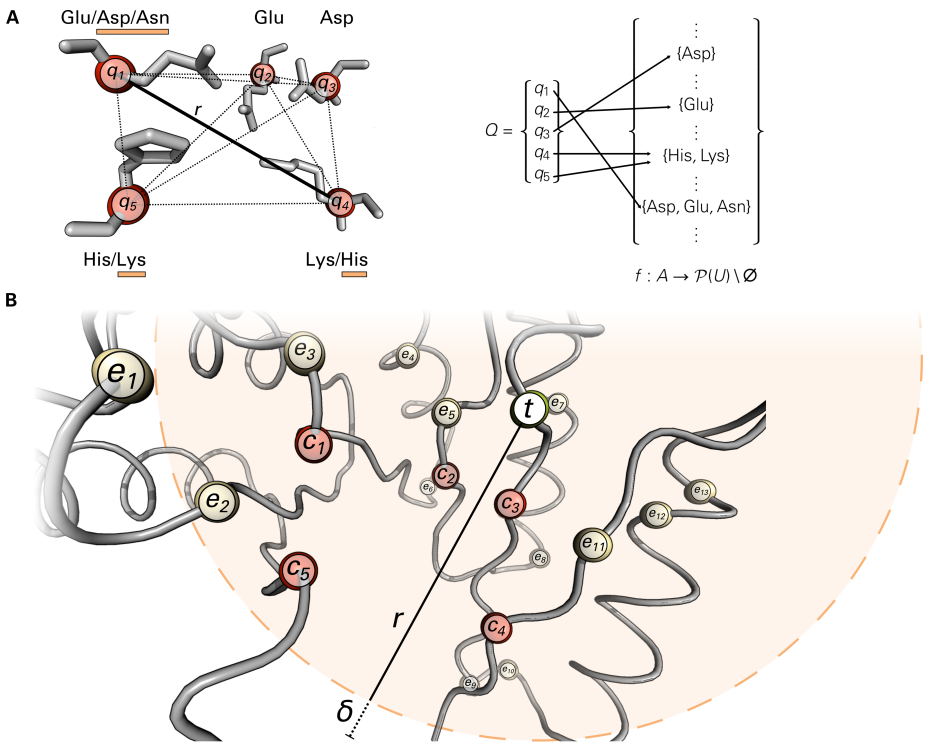


Figure 6.2.: Illustration of the template-based combinatorial search approach used by Fit3D. (**A**) Determination of maximal spatial extent r and the mapping $f(q)\ \forall\ q \in Q$ for the ES motif [33]. PSEs are marked with an orange box. (**B**) Iterative search in a target structure T, the extraction of a local environment $E^t = \{e_1, ..., e_{13}\}$ around amino acid t, and the determination of match candidates $C = \{c_1, c_2, \ldots, c_5\}$.

Iterative Detection The detection process is an iteration over all amino acids $t \in T$ in the target structure T (Algorithm 1, Line 3). If $g(t)$ is a subset of at least one set of allowed amino acid labels $f(q)$ for each $q \in Q$, the local environment E^t around t is extracted within the radius $r + \delta$ (see Figure 6.2B and Algorithm 1, Line 5). The parameter δ is the distance tolerance threshold with a default value of 1.00 Å. The local environment extraction is based on the coordinates of alpha carbon atoms. If pairwise distance filtering of E^t is enabled, only pairs of $(e_v, e_w) \in E^t, v \neq w$ are kept where a corresponding pair $(q_x, q_y) \in Q, x \neq y$ exists such that residue labels are compatible and the distance is similar: $\|e_v - e_w\| \leq \|q_x - q_y\| + \delta \ \wedge\ g(e_{v,w}) \subset f(q_{x,y})$. Subsequently, all k-sized match candidates $\mathcal{C} = \binom{E^t}{k}$ of the local environment are determined. Because only a limited number of these combinations is valid, pruning of the combinatorial space is conducted as early as possible (see Figure 6.3).

Assessment of Match Candidates The optimal superimposition out of $k!$ possible superimpositions is determined for each valid combination $C \in \mathcal{C}$ where the compatibility of amino acid labels is given: $\forall\ q \in Q\ \exists\ c \in C : g(c) \subset f(q)$ (see yellow spheres in Figure 6.2B). Again, the number of superimposition can be drastically reduced by considering only permutations which are compatible to the template motif regarding their residue labels. For each valid permutation the geometric dissimilarity $d = \mathrm{RMSD}(C, Q)$ is determined. The optimal superimposition out of all valid superimpositions $s \in S(C)$ minimizes d: $\min_{s \in S(C)} \mathrm{RMSD}(s, Q)$. Only if $d \leq \varepsilon$, i.e. the geometric dissimilarity is below the desired RMSD upper bound, the candidate set is considered to be a match. Illustratively, in Figure 6.2B this is the case for the candidates $C = \{c_1, c_2, \ldots, c_5\}$ shown in red.

Algorithm 1: Pseudocode of the template-based Fit3D algorithm.

Input: template structural motif Q of size k, target structure T, distance tolerance δ, RMSD upper bound ε

Output: $C \subset \binom{T}{k}$ with $g(c) \subset f(q),\ \forall\ c \in C,\ \forall\ q \in Q$ and $\mathrm{RMSD}(C, Q) \leq \varepsilon$

```
1  begin
2      r ← max(‖q_i − q_j‖ : q_{i,j} ∈ Q, i ≠ j)      // determine template motif extent
3      for t ∈ T do
4          if g(t) ⊂ f(q) ∀ q ∈ Q then
5              E^t ← {t′ ∈ T\{t} : ‖t − t′‖ ≤ r + δ}      // compose local environment
6              if filter then
7                  E^t ← E^t ∩ {‖e_v − e_w‖ ≤ ‖q_x − q_y‖ + δ ∧ g(e_{v,w}) ⊂ f(q_{x,y})}
8              end
9              if |E^t| < k then
10                 continue
11             end
12             for C ∈ (E^t choose k) do
13                 if ∀ q ∈ Q ∃ c ∈ C : g(c) ⊂ f(q) then
14                     d ← RMSD(C, Q)      // calculate RMSD by superimposition
15                     if d ≤ ε then
16                         accept C
17                     end
18                 end
19             end
20         end
21     end
22 end
```

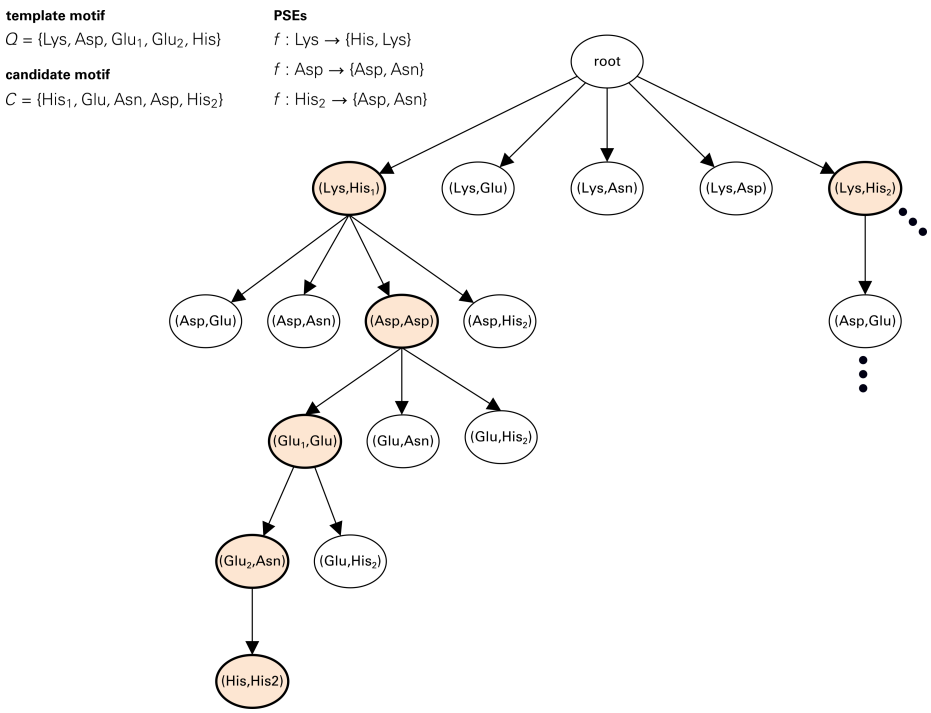


Figure 6.3.: The generation of valid candidates used for alignment against the template motif. For a given template motif Q of size k and a match candidate C a search tree is constructed by pairing the two sets. For each tuple in the search tree, it is determined whether the pairing is valid, i.e. the result of the mapping function f contains the label of the candidate residue. Only if the label is identical or PSEs are valid, the search tree is extended. A valid alignment between Q and C corresponds to every path of the search tree with length k (orange shaded and emphasized nodes).

6.1.3. BENCHMARK AND VALIDATION

In the following, the template-based detection with Fit3D is tested in respect to the runtime of real-world applications and the general performance on selected benchmark datasets. These datasets focus on the adequate computational representation of structural motifs and show how the key features of Fit3D allow an increase in specificity and sensitivity of the method.

TIME COMPLEXITY

The spatial extent r of the template motif can be found in time $\mathcal{O}(k^2)$ by computing k^2 distances for all pairs of alpha carbon atoms in Q. Furthermore, the match candidates $\mathcal{C} = \binom{E^t}{k}$ for a local environment can be calculated within $\mathcal{O}(\binom{l}{k-1})$ time with $|E^t| = l$. Due to the fact that the current target amino acid is stored and only $k - 1$ amino acids in E^t have to be considered in order to find a valid combination. To find the optimal superimposition in the worst case $k!$ calculations have to be performed. In fact only amino acids with compatible labels are aligned which reduces the number of necessary alignment steps, and thus the runtime, significantly. The algorithm complexity has to be defined in dependence of local filtering and differs for searches with and without local filtering. In general, the application of filtering speeds up runtime substantially, especially for motifs with large spatial extent r. The worst time complexity of the algorithm without pairwise distance filtering the local environment is $\mathcal{O}(k^2 + n\binom{l}{k-1}k!\Theta(k))$ where $\Theta(k)$ is the time required to evaluate geometric constraints (the superimposition) of a set of k amino acids. Otherwise, if filtering is enabled, it takes additionally $\mathcal{O}(l^2)$ time to compute pairwise distances of the environment, but afterwards $\Phi(l)$ less combinations have to be computed, where $\Phi(l)$ discards amino

acids not fulfilling the pairwise distance constraints: $\mathcal{O}(k^2 + n(l^2 + \binom{l-\Phi(l)}{k-1})k!\Theta(k))$. This is especially important if the template motif has a high spatial extent which leads to bigger local environments with more residues to consider. Additionally, it has to be admitted that the complexity of the Fit3D motif matching algorithm strongly depends on the density and size of the local environment. For target amino acids t buried in the protein, l tends to be bigger than for target amino acids at the protein surface and vice versa. In the worst case l could be as large as n. However, this is virtually impossible if Fit3D is used appropriately because the definition of small and locally occurring structural motif (see Chapter 3, Definition 3.1) implies that $l \ll n$.

RUNTIME ANALYSIS

The Fit3D algorithm was investigated regarding its runtime for real-world applications. This includes the parsing of macromolecular structures as well as the subsequent detection process. Based on the capability of Fit3D to benefit from the Macromolecular Transmission Format (MMTF), a new transmission format for macromolecular data [201], the first benchmark aimed at testing the influence of parsing structures with MMTF.

Serine Proteases Catalytic Triad The catalytic triad of serine proteases consisting of histidine, aspartic acid, and serine [2] was detected in different-sized non-redundant subsets of the PDB. Figure 6.4A shows the result of this benchmark case. The average runtime of Fit3D amounts to a value of 46 ms/structure if structures are parsed in the conventional PDB format and 3 ms/structure if MMTF is used. This corresponds to a 15-fold speedup for MMTF. Beside the substantial increase in processing time, MMTF is memory efficient as is stores only a compressed and reduced version of the structure [201]. For the investigated database sizes, the runtime stays almost constant for MMTF, while usage of PDB parsing shows a steady linear increase in runtime. For the particular problem MMTF seems to scale much better than the conventional PDB format. Beside this, MMTF is not limited to a maximal number of 10,000 atoms – as it is the case for the PDB format – and can handle large macromolecular structures such as ribosomal subunits.

Catalytic Site Atlas Motifs A second benchmark was conducted to investigate the scalability of Fit3D for the detection of structural motifs with different properties. For this analysis, a non-redundant subset of all structural motifs with a size between two and six residues, annotated as active sites in the CSA [124], was used. The target dataset contained a PDB subset of 100 non-redundant structures and was parsed in PDB format. Figure 6.4B shows the results of this benchmark. In total, 283 structural motifs were tested for their runtime with Fit3D in five independent runs. The results indicate the spatial extent of the template motif to be the limiting factor for processing time. This coincides with the complexity of the Fit3D algorithm, which strongly depends on the size of the extracted local environments as elucidated beforehand. The runtime seems to exhibit an exponential increase for structural motifs of at least five residues and a maximal spatial extent over $\approx$14 Å. However, the template-based detection for most of the tested structural motifs can be computed in reasonable time. The processing of 95% of the CSA-derived structural motifs takes less than 136 ms/structure. The active site motif of a phosphoserine phosphatase (PDB:1l7n) showed a very high average processing time of 101,044 ± 2,597 ms for the detection in 100 structures. Closer investigation of this motif did not show any salience in respect to its spatial extent. It is likely that this motif has a very generic residue composition resulting in many candidates that have to be checked for compatibility.

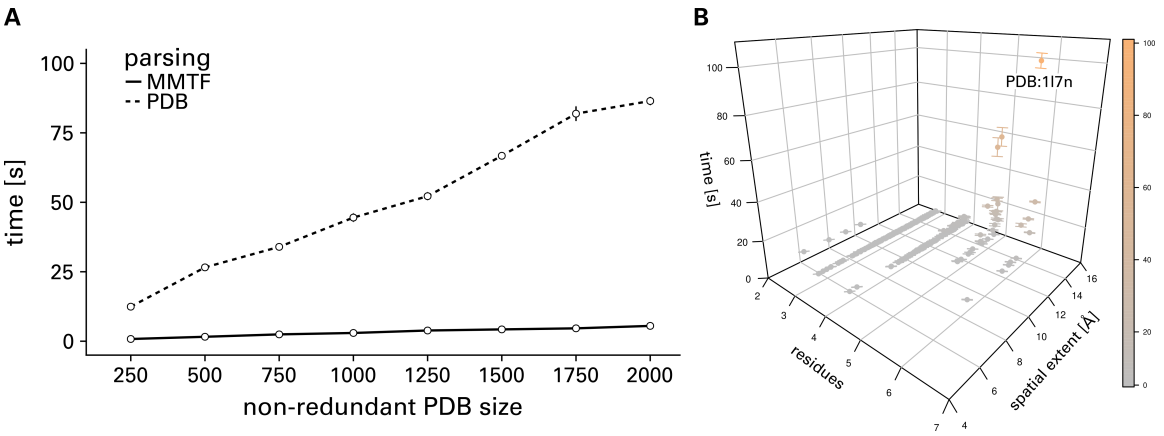


Figure 6.4.: Runtime benchmarks of template-based structural motif detection with Fit3D. (**A**) The runtime of the Fit3D algorithm for the detection of a single structural motif in seconds depending on the parsing strategy: PDB format or MMTF [201]. The serine protease catalytic triad was extracted from structure PDB:1gl0 and served as a template motif for the detection in non-redundant datasets of protein structures of different size. Error bars indicate the standard deviation of five independent runs. (**B**) Runtime of Fit3D in dependence of 283 motifs of different size and spatial extent derived from the CSA [124]. The runtime in seconds is plotted against the size and the spatial extent of the motifs. Error bars indicate the 95% confidence interval of five independent runs.

VALIDATION

The validation of substructure matching algorithms based on ES structures and CSA-derived motifs is the de facto standard and was successfully applied several times [43, 137, 142, 161]. Hence, the following section presents the results for a validation of Fit3D based on two datasets: structures of the ES derived from the SFLD [202] and structures of the Nitric Oxide Synthase (NOS) family derived from the CSA [124]. The experiment for both of these benchmarks consists of recovering true positive matches from a background dataset. Based on this setup, the sensitivity and specificity of Fit3D can be assessed. Only significant matches with p-value<0.001 according to the statistical model of FOFANOV ET AL. are considered for the assessment [161]. To highlight the importance of adequate computational representation of structural motifs, all benchmarks were carried out for two representation schemes: alpha carbon atom and all-atom.

Nitric Oxide Synthase Figure 6.5A shows the results of a benchmark of the Fit3D algorithm. The benchmark focused on the detection of a specific structural motif in NOS, a class of enzymes that synthesize nitric oxide from L-arginine [203]. The template structural motif was defined as in [43] and contained cysteine, arginine, tryptophan, and glutamic acid (see Figure 6.5B). The accuracy ($\frac{TP+TN}{TP+FP+FN+TN}$) differs only marginally between the two tested representation types. It experiences a minor increase from $\frac{124+38,292}{124+1,300+142+38,292}$=0.9638 to $\frac{262+39,505}{262+87+4+39,505}$=0.9977 if all atoms of the template motif are considered. However, the test dataset is very unbalanced. It contains by two orders of magnitude more negative instances (39,590) than instances actually belonging to the NOS family (266). Hence, accuracy is an inadequate measurement for actual performance. This is reflected by the stark increase in sensitivity (the true positive rate, $\frac{TP}{TP+FN}$) of 51.88%, from $\frac{124}{124+142}$=0.4662 to $\frac{262}{262+4}$=0.9850, if all atoms are considered. Consequently, all-atom representation avoids the reporting of many false negative matches. In terms of specificity (the true negative rate, $\frac{TN}{TN+FP}$) there is only a minor difference of 3.10% between alpha carbon ($\frac{38,293}{38,292+1300}$=0.9672) and all-atom ($\frac{39,505}{39,505+87}$=0.9978) representation.

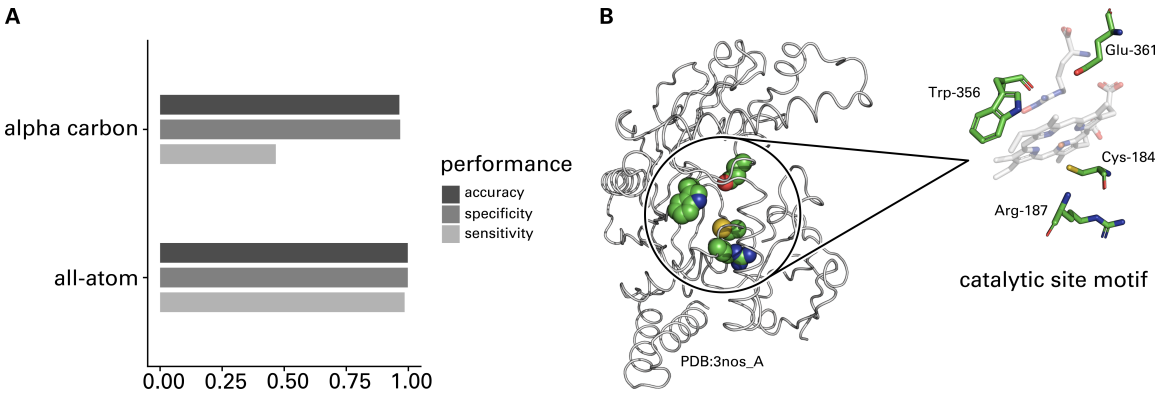


Figure 6.5.: Results of a template-based detection of the NOS catalytic site motif. (**A**) The performance of Fit3D in respect to different computational representations of structural motifs. (**B**) An example structure of the NOS protein family (PDB:3nos chain A) shown in wire representation. The structural motif in the active site as annotated in the CSA [124] is highlighted. It consists of cysteine 184, arginine 187, tryptophan 356, and glutamic acid 361. The both ligands, arginine and heme, are shown as translucent sticks.

Enolase Superfamily Figure 6.6A shows the results of a benchmark of the Fit3D algorithm to detect a superfamily-representing structural motif. The ES is a class of enzymes that catalyze a variety of biochemical reactions. However, they share a common reaction: the abstraction of a α-proton of carboxylic acids. A template structural motif for ES, representative for all superfamily members, was defined as described in [33]. The motif is variable in amino composition and thus requires the use of PSEs. It consists of lysine/histidine, aspartic acid, glutamic acid, glutamic acid/aspartic acid/asparagine, and histidine/lysine (see Figure 6.6B). If Fit3D is run with single-point representation of the template structural motif via its alpha carbon atoms, the overall performance is considerably lower in comparison with all-atom representation. The accuracy for alpha carbon representation is $\frac{61+28,833}{61+10,759+12+28,833}=0.7285$ compared to $\frac{68+39,062}{68+530+5+39,062}=0.9865$ for all-atom representation. This corresponds to an increase of 25.81% in accuracy if all atoms of the template structural motif are considered for the detection process. The values for sensitivity increase from $\frac{61}{61+12}=0.8356$ to $\frac{68}{68+5}=0.9315$ if all-atom representation is enabled. A remarkable increase in specificity of 25.84% ($\frac{28,833}{28,833+10,759}=0.7283$ to $\frac{39,062}{39,062+530}=0.9866$) is also evident if all atoms are considered for the detection. In general, the capability of Fit3D to make use of full atomic resolution allows for both less false negative and less false positive matches. The NOS dataset contained 73 positive and 39,518 negative instances.

6.1.4. IMPLEMENTATION

To increase usability of the developed method, Fit3D is provided as standalone command line implementation, web server, and API version. The three implementations address different user groups. While the non-expert users benefit from the easy-to-use web interface, advanced users might consider integrating Fit3D into custom workflows via its API version. All versions are open source and were implemented in Java in order to enable platform independence.

Command Line Implementation The command line implementation[6] of the algorithm was designed to be simple in usage and flexible in application. It offers a command line interface and allows the easy adjustment of algorithmic parameters; the full-fledged features of Fit3D are accessible via a range of different command line flags (see Table B.1). Statistical significance estimation of matches according to [161] or [162] is supported. The

[6]available at: github.com/fkaiserbio/fit3d/releases

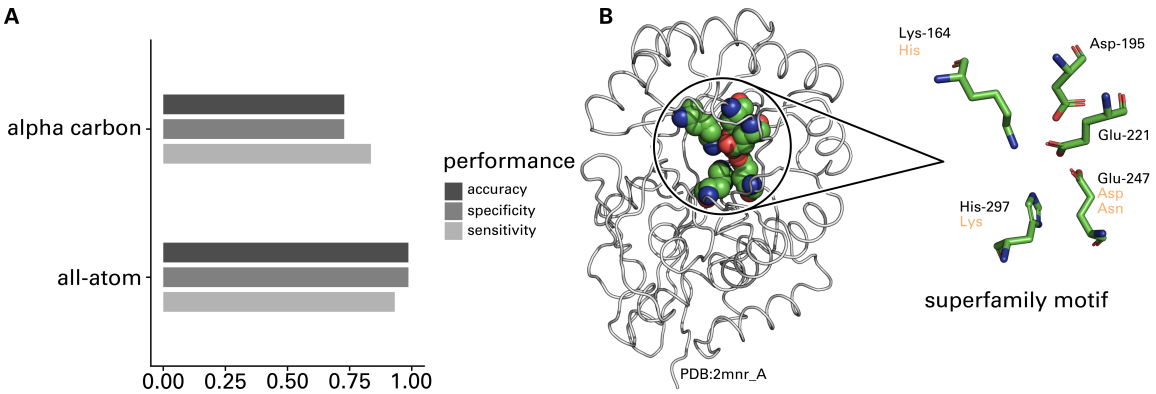


Figure 6.6.: Results of a template-based detection of a structural motif, representative for the ES [33]. (**A**) The performance of Fit3D in respect to different computational representations of structural motifs. (**B**) The structure of origin of the template motif, a mandelate racemase from *Pseudomonas putida*. The structural motif as described in [33] is highlighted and consists of the residues lysine 164, asparagine 195, glutamic acid 221, glutamic acid 247, and histidine 297. The residues at certain positions are annotated with PSEs (orange labels).

calculations are automatically run in parallel, unless explicitly deactivated, on all available processing cores of the machine. The basic output of a template-based detection is a report of all matches in comma-separated values (CSV) format and, if activated, PDB structures of matches aligned to the template motif for an easy visual assessment of the results. For details on the usage please refer to Appendix B.1.

Web Server In addition to the command line version of Fit3D, an intuitive web server was designed[7]. It offers an "all-in-one" solution and covers the whole process of template-based structural motif detection from template definition (Figure 6.7A) and subsequent database screening up to the analysis and visualization of the obtained matches (Figure 6.7B). The mapping of UniProt [175], Pfam [124], and EC [41] annotations allows a fast identification of relevant matches. Furthermore, an interactive visualization [204] of structures is implemented. Different visualization types can be picked: alignment of all matches (or a single match) against the template motif and the global superimposition of two structures based on the found match. All results are provided as individual files or an archive file, which bundles all aligned matches in PDB format plus a summary file in CSV format. The distribution of the RMSD for all matches is visualized and available for download, which can be an important signature pattern depending on individual characteristics of the template motif. The Fit3D web server is optimized for the detection of small structural motifs up to a size of five residues. Recall of the results is possible within 72 hours after calculation via an individual link, which is sent to the user by email. The Fit3D web server is based on the SiNGA API [170] and supports parsing of MMTF. Details on the usage are given in Appendix B.2.

API Version In addition to the command line and web server versions of Fit3D, an API version is available. This version is highly customizable and allows the integration into specialized workflows for the expert user. It is an integral part of the SiNGA framework [170] that features many tools for the analysis of macromolecular data. Hence, no additional dependencies are required. SiNGA is deployed to the Maven Central Repository[8] and requires the Java Development Kit 1.8 or later. Please refer to the documentation of SiNGA[9]

[7] available at: biosciences.hs-mittweida.de/fit3d/
[8] available at: mvnrepository.com/artifact/de.bioforscher.singa
[9] available at: github.com/cleberecht/singa/wiki/Structure-Alignments-(Chemistry)

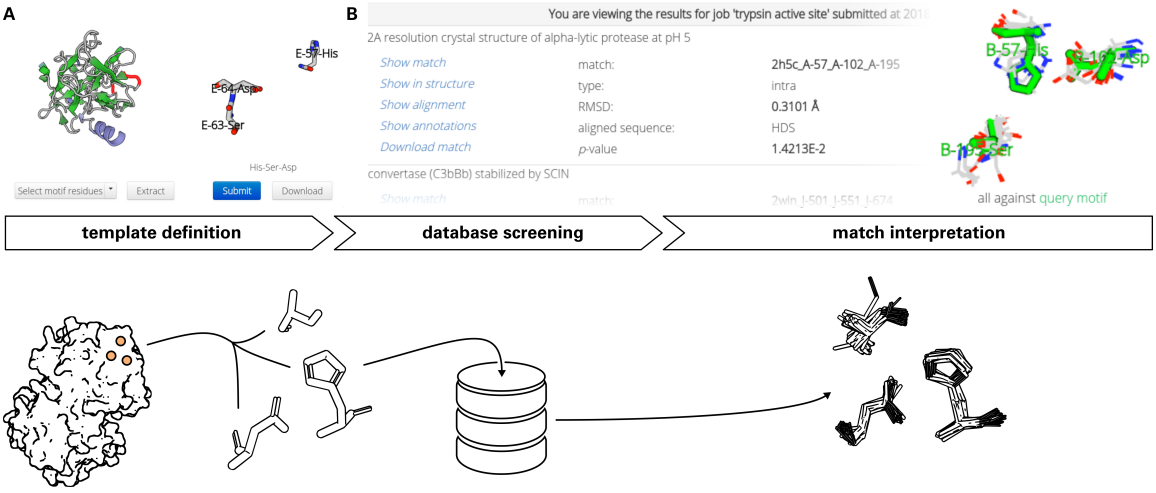


Figure 6.7.: The template-based structural motif detection workflow with the Fit3D web server. (**A**) Based on structure data, uploaded by the user or retrieved from the PDB, the motif extraction wizard allows the definition of a template motif. Subsequently, the detection of the template motif in predefined (or user-provided) datasets of non-redundant structures is available. (**B**) The comprehensive output of the web application features detailed results for each obtained match. The distribution of all RMSD values is visualized and available for download. Different interactive visualization options can be selected and all results can be downloaded as a single archive file that contains results in CSV format and structures in PDB format.

or to Appendix B.3 on how to run a search with Fit3D at the API level.

6.2. TEMPLATE-FREE STRUCTURAL MOTIF DETECTION

The capability of Fit3D to detect matches of structural motifs based on a given template is complemented by a template-free detection engine. This section introduces template-free structural motif detection and presents an adaption of a popular data mining technique to solve the problem. The developed algorithm can be used to investigate macromolecular structures but the basic concept is applicable to any kind of labeled spatial data.

6.2.1. PROBLEM ORIGIN

In contrast to template-based structural motif detection, where a template motif is available *a priori*, template-free detection is a pattern discovery problem in the three-dimensional space. It can be addressed using diverse approaches from data mining, such as frequent subgraph mining [153]. In general, the task is to identify subsets of multiple point clouds in the three-dimensional space that are geometrically similar (Figure 6.8). The mining of spatial data has been applied on meteorological, disease outbreak, or medical imaging data [148]. Protein structure data is labeled – each amino acid has an associated residue type – and can thus be analyzed using so-called itemset mining. A definition of the term itemset in the context of structural motif detection is given in Definition 6.1.

Definition 6.1 (Itemset)**.** An itemset is a set of labeled items. Items in the set must be unique regarding their labels; repetition of items with identical labels is prohibited. In the context of structural motif detection, an itemset is equivalent to a set of different amino acids. A frequent itemset is repetitive in the analyzed data and can thus be considered to be the equivalent of a conserved structural motif (see Chapter 3, Definition 3.2).

Classical itemset mining originated from data mining ("market basket analysis") and has been used for different applications such as text mining, time-series, graph, or spatial data

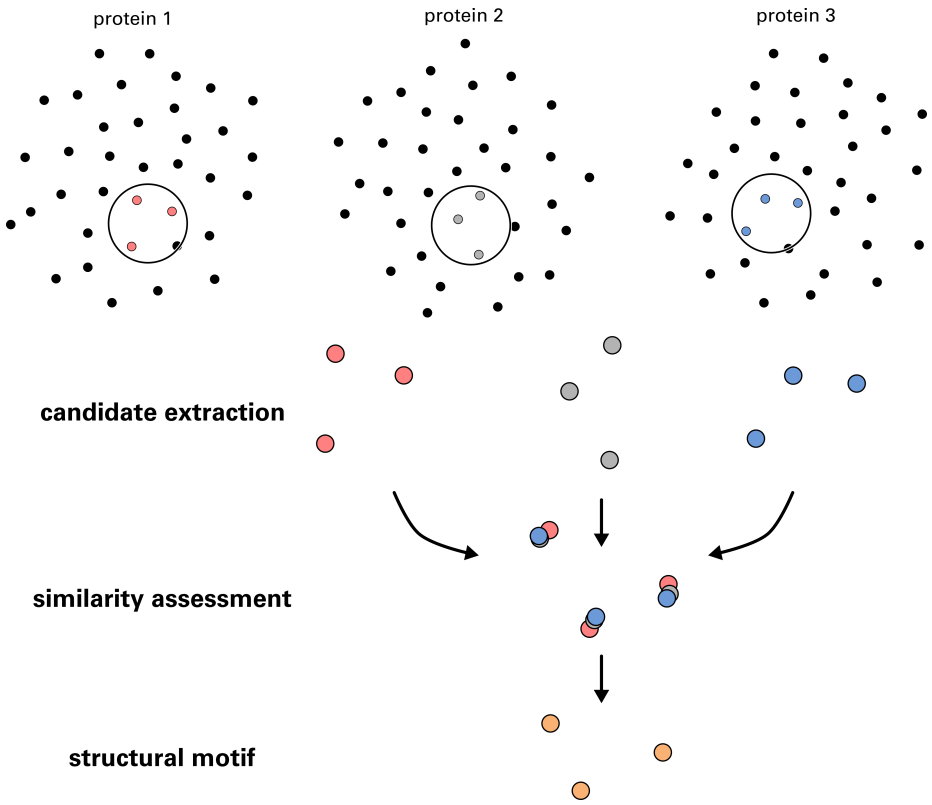


Figure 6.8.: Template-free structural motif detection can be seen as pattern discovery in the three-dimensional space. Protein structures are represented as points. In order to detect common structural motifs in a given set of protein structures candidates are extracted and subsequently assessed for similarity. If similar patterns are frequent in the dataset of protein structures, they correspond to a structural motif.

analysis [148]. The generalized aim of itemset mining is to determine associations between items, which can be expressed by different measurements, the so-called metrics. One basic metric is the support – a measurement that describes the relative occurrence of itemsets in the data.

Itemset mining was already applied successfully to protein data, e.g. to identify binding motifs for transcription factors or splicing patterns [149]. However, in 2014 ZHOU ET AL. were the first to suggest using frequent itemset mining to spot interesting biological patterns in protein structure data without any abstraction of spatial data to graphs, distance matrices, or structural features [150]. They introduced the concept of cohesion to avoid the explicit restriction of distances between items during the mining process and to discover patterns in spatial proximity, i.e. cohesive patterns. The method was applied to different protein families and revealed cohesive patterns that span large distances at sequence level but are brought to proximity in tertiary structure by protein folding. The authors suggest that these patterns play a role for the specific or overall structure of the protein [150]. In a subsequent study authored by the same group, their methods were extensively applied to a PDB-wide dataset in order to identify cohesive patterns not linked to any concrete fold or function. These patterns occur mainly beyond annotated protein domains [151]. Additionally, the method was used to mine specific cohesive patterns, which correlate with optimal growth temperature of different prokaryotic species, and to identify preferential contacts in DNA-binding proteins [151]. This highlights the dualistic character of structural motifs: family-specific or function-related, as well as unspecific and ubiquitous, with the latter to be seen as common molecular building blocks for structure stabilization.

6.2.2. ALGORITHM

The following section presents an algorithm for the unsupervised and template-free detection of structural motifs. The algorithm detects motifs of high geometric similarity in a dataset of protein structures, which is usually the case for functionally relevant structural motifs as shown for aaRSs (Chapter 5). The method allows pinpointing structural motifs in a high-throughput manner. It is not limited to proteins, but can also be applied to any type of macromolecular structures, such as DNA, RNA, or labeled spatial data in general. The fundamental basis of the proposed method is the concept of cohesion that was introduced by ZHOU ET AL. in 2014 [150]. In the following, this concept is extended by defining novel metrics to optimize the mining results and to obtain structural motifs with similar geometry.

Assumptions Itemset mining can be adapted for protein structure data as follows. Consider $U = \{\text{Ala}, \text{Arg}, \text{Asn}, ..., \text{Val}\}$ (the universe of items) to be the three-letter code labels of the 20 canonical amino acids, $\mathcal{T} = \{T_1, ..., T_n\}$ (the transactional data) a database of protein structures, and $T = (t_1, ..., t_m)$ a single protein within this database. Each amino acid $t = (u, v)$ of the protein is composed of a label $u \in U$ and coordinate information $v \in \mathbb{R}^3$. Furthermore, T defines an intrinsic order of elements. The mapping function $L : T \mapsto U$ maps amino acids to their labels. The itemset $I = \{u_1, ..., u_k\}$ denotes a k-itemset, e.g. $I = \{\text{Ile}, \text{Val}\}$ is a 2-itemset of isoleucine and valine. Frequent itemsets $\mathcal{I}$ are all k-itemsets out of all possible $2^{|U|} - 1$ nonempty itemsets with $1 \leq k \leq |U|$, such that each itemset obeys a cutoff ε depending on the respective evaluation metric function $E : I \mapsto \mathbb{R}$:

$$\forall I \in \mathcal{I} : E(I) \begin{cases} \leq \varepsilon & \text{if } E(I) \text{ should be minimized,} \\ \geq \varepsilon & \text{if } E(I) \text{ should be maximized.} \end{cases} \tag{6.1}$$

Candidate Generation As shown in Figure 6.8, candidates for structural motifs have to be extracted from protein structures in the database. In order to generate these candidates, which are then evaluated in respect to certain metrics and assessed for similarity, the popular Apriori algorithm was used [205]. Even though more time-efficient algorithms for candidate generation are available nowadays [149, 206], Apriori was chosen due to its simplicity and because most of the computation time is dedicated to metric evaluation and not candidate generation. In each round the Apriori algorithm extends itemsets from the previous round with one item at the time. For example, the candidates of the first round are 1-itemsets representing the 20 canonical amino acids: {Ala}, {Arg}, {Asn}, ..., {Val}. These candidates are then evaluated against the database $\mathcal{T}$ and only retained if they pass all tests (the evaluation metric functions). Candidates that do not pass the tests are pruned. For example, if all k-itemsets pass, the next candidates consist of all possible $k\frac{k-1}{2}$ extensions of k-itemsets. In the case of amino acids, candidates of the second round would be 190 2-itemsets composed of pairs of the 20 canonical amino acids, e.g. {Ala, Asp}, {Ala, Cys}, or {Asp, Cys}. Hence, itemset mining is the iterative process of the following steps:

- generation of k-sized candidates of current round k,
- calculation of evaluation metric functions,
- pruning of generated candidates that constitute the input of the next round $k+1$, and
- termination if all candidates were pruned.

The input selection for the next round follows the so-called downwards closure property: every subset of a frequent itemset is also frequent [148]. The mining process converges

and the algorithm terminates if all candidates for the next round input $k+1$ were pruned. In the following, fundamental and extended evaluation metrics are elucidated.

EVALUATION METRICS

Itemsets can be evaluated using different metrics in order to decide whether they are within the scope of interest and part of the next candidate generation round. Simple evaluation metrics operate on the database $\mathcal{T}$, not taking into account any spatial information associated with database entries and not requiring any annotated metadata.

Support The support of an itemset is defined as the fraction of the number $N(I)$ of proteins T that contain I as a subset of their labels $L(T)$ with respect to the size of the dataset:

$$\text{support}(I) = \frac{N(I)}{|\mathcal{T}|} \tag{6.2}$$

where

$$N(I) = \sum_{T \in \mathcal{T}} \begin{cases} 1 & \text{if } I \subset L(T), \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

The support should be maximized if one wants to identify itemsets which occur frequently across the database.

EXTRACTION METRICS

Extraction metrics are evaluation metrics capable of extracting concrete observations $\mathcal{W}(I)$ of a k-itemset I from the database $\mathcal{T}$. The labels $L(W) = \{u_1, ..., u_k\}$ of each itemset observation $W \in \mathcal{W}$ are identical to I:

$$\mathcal{W}(I) : \{W \in \binom{T}{k} : L(W) = I\} \; \forall \; T \in \mathcal{T}. \tag{6.4}$$

Hence, itemset observations correspond to all possible combinations of k-sized subsets per database entry T with the same labels as itemset I. The function $O : \mathcal{W} \mapsto \mathcal{T}$ maps each itemset observation to its database entry of origin.

Cohesion ZHOU ET AL. introduced the cohesion measurement to evaluate itemset candidates [150]. Cohesion uses the coordinate information associated with each amino acid in the protein to detect spatially-cohesive itemsets, i.e. amino acids that occur in spatial proximity in the three-dimensional structure of the protein. This is one of the previously defined criteria for structural motifs (Chapter 3, Definition 3.1). In order to calculate cohesion, one has to find the ball with the smallest radius (the smallest enclosing ball) that contains a set of amino acids. These amino acids are represented by points in the three-dimensional space, e.g. by their alpha carbon atoms or geometric centers. For a given set of points the smallest enclosing ball was shown to exist and to be unique [207]. Hence, to determine the cohesion for a given itemset I, the enclosing ball with the smallest radius per database entry T has to be found, considering each possible itemset observation $\mathcal{W}(I)$. In order to calculate cohesion by determining the smallest enclosing ball the heuristic VertexAll algorithm [150] was implemented. The function $R : \mathcal{W} \mapsto \mathbb{R}_{>0}$ determines the enclosing ball radius for a given itemset observation based on the associated coordinate information. Cohesion can now be defined as the averaged sum of all enclosing ball radii of the itemset observations $\mathcal{W}(I)$ that are found in the database, whereby only the itemset observation with the smallest enclosing ball radius per database entry T is considered:

$$\text{cohesion}(I) = \frac{1}{N(I)} \sum_{W \in \mathcal{W}} R_{\min}(W) \tag{6.5}$$

with

$$R_{\min}(W) = \begin{cases} R(W) & \text{if } R(W) = \min\limits_{W' \subset O(W)} R(W'), \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

Because the cohesion metric considers only the itemset observation with the smallest enclosing ball radius, cohesion is suitable to discover structural motifs that occur only once in a protein structure, such as catalytic sites [122]. While mining for structural motifs the cohesion should be minimized.

Adherence In order to overcome the limitation of cohesion where only one itemset observation per structure is considered, a new metric is introduced. This metric is an extension of the cohesion concept, not only considering itemset observations with the smallest enclosing ball radius per database entry, but rather all itemset observations $\mathcal{Y} \subset \mathcal{W}(I)$ with their enclosing ball radii within a certain threshold δ with respect to a desired radius r:

$$\mathcal{Y} : \{W \in \mathcal{W}(I) : r - \delta \leq R(W) \leq r + \delta\}. \tag{6.7}$$

The measurement of adherence can now be defined as the standard deviation of the enclosing ball radii:

$$\text{adherence}(I) = \sqrt{\frac{1}{|\mathcal{Y}|} \sum_{Y \in \mathcal{Y}} (R(Y) - \mu)^2} \tag{6.8}$$

where

$$\mu = \frac{1}{|\mathcal{Y}|} \sum_{Y \in \mathcal{Y}} R(Y). \tag{6.9}$$

Because all itemset observations with a smallest enclosing ball radius within $r \pm \delta$ are considered, the adherence metric is suitable to discover structural motifs that are recurrent within a single protein structure. This is important in order to find allosteric, protein-protein interaction, or ion-binding sites [122]. As the adherence represents the standard deviation of a desired radius r it should be minimized during the mining process. For a visual representation and comparison of cohesion and adherence refer to Figure 6.9.

EXTRACTION-DEPENDENT METRICS

The concepts of cohesion and adherence can be exploited to extract concrete observations of itemsets with a desired cohesion or adherence from the data. Subsequently, these observation can be further evaluated regarding other metrics. This paves the way to apply diverse additional metrics to assess the relevance of itemsets regarding special requirements, e.g. high geometric similarity or separation at the sequence level. These metrics depend on the availability of concrete itemset observations and are consequently called extraction-dependent metrics.

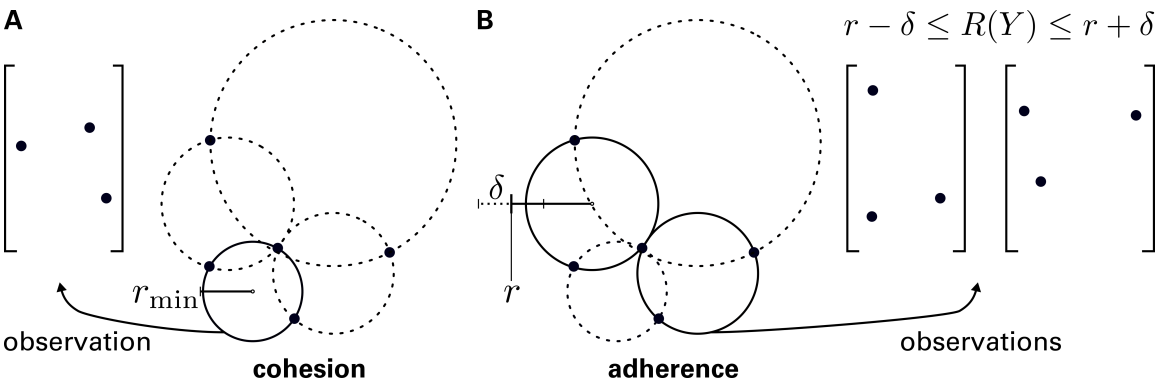


Figure 6.9.: Comparison of extraction metrics. (A) Extraction of itemset observations with cohesion. The cohesion extraction metric extracts exactly one itemset observation for each database entry. Ideally, the observation with the smallest enclosing ball radius r_{min} is chosen (solid circle), determined by the heuristic VertexAll algorithm [150]. (B) Extraction of itemset observations with adherence. The adherence extraction metric extracts itemset observations with enclosing ball radii close to the desired radius r. Hence, for each data point several observations can be extracted (solid circles).

Consensus In order to precisely identify geometrically conserved structural motifs, the consensus metric is introduced, which uses the coordinate information $v \in \mathbb{R}^3$ associated with each amino acid of a protein. The geometric similarity of structural motifs is another important aspect (see Chapter 3, Definition 3.2) to infer functional relevance. A flexible computational representation of amino acids by arbitrary atoms, the last heavy side chain atom, the centroid of all atoms, or the centroid of side chain atoms is implemented. To assess the geometric similarity of itemset observations, an adaption of the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [208] is proposed, a hierarchical clustering algorithm that operates on a distance matrix. UPGMA is used in bioinformatics to construct phylogenetic trees of sequence data or to create guide trees for MSAs [209]. The adapted UPGMA algorithm is shown in Algorithm 2. It uses a dissimilarity matrix of all pairwise distances (RMSD values) of itemset observations $\mathcal{W}(I)$ as input. Subsequently, UPGMA locates the pair with the smallest distance value. This corresponds to the identification of the geometrically most similar pair $(W_i, W_j)_{\text{min}}$ of itemset observations, i.e the pair with the lowest RMSD value (Algorithm 2, Line 12). Instead of joining lineages as in the original UPGMA algorithm, the adaption now averages the coordinates of the two closest itemset observations (a new cluster is formed) and creates a consensus observation $\overline{W}$. The closest pair is now removed from the itemset observations $\mathcal{W}(I)$ and the consensus observation $\overline{W}$ is added (Algorithm 2, Line 15 and Line 16). The distance matrix, reduced in size by one row and one column, is recalculated. Again, the closest pair is identified and the procedure is repeated until all itemset observations are processed and the cardinality of $\mathcal{W}(I)$ is one. To obtain the consensus score of one itemset

$$\sum_{n=1}^{|\mathcal{W}(I)|} n\frac{n-1}{2} \tag{6.10}$$

alignments have to be computed in total, which is the most time-consuming task of the presented workflow. For each alignment the optimal superimposition [136] has to be calculated, resulting in $4{,}995 \cdot 10^8$ alignment calculations for an itemset with 1,000 observations. Thus, a sampling-based scheme is applied that restricts the number of itemset observations which should be considered. The algorithm randomly selects itemset observations such that $|\mathcal{W}(I)| \leq 1,000$. Therefore, up to 1,000 structures are considered for consensus calculation if the cohesion metric is used, because one itemset observation is extracted per structure. If the adherence metric is used, the number of considered structures depends

on the amount of itemset observations found within the enclosing ball radius $r \pm \delta$.

The consensus measurement is then defined as the sum of all RMSD values of superimposed pairs of itemset observations, normalized by the number of coordinate averaging steps taken during the UPGMA clustering procedure (Algorithm 2, Line 14) and the size k of the itemset:

$$\text{consensus}(I) = \frac{1}{k(|\mathcal{W}(I)| - 1)} \sum_{\substack{W_i, W_j \in \mathcal{W}(I) \\ i \neq j}} \text{RMSD}(W_i', W_j). \tag{6.11}$$

Due to the UPGMA-like approach a hierarchical clustering is obtained; it is possible to construct a tree that represents the geometric relationship between each observation $W \in \mathcal{W}(I)$ of itemset I. Tree leaves are itemset observations $\mathcal{W}(I)$ whereas nodes are consensus representations $\overline{\mathcal{W}}(I)$ originated from averaging coordinates of superimposed atom pairs. The branch length b is equidistant between two itemset observations (W_i, W_j) and their consensus observation $\overline{W_{i,j}}$ and defined as

$$b(W_i, \overline{W_{i,j}}) = b(W_j, \overline{W_{i,j}}) = \frac{\text{RMSD}(W_i', W_j)}{2}. \tag{6.12}$$

This allows for the clustering of itemset observations according to their structural similarity by splitting the obtained tree at a certain depth λ. Beside this, an abstract representation of each cluster is given by its top-level consensus, which can be seen as a template structural motif representative for the cluster. The consensus metric allows to detect geometrically conserved structural motifs that are prevalent in the set of target protein structures. It should be noted that this approach is applicable to discover different structural or physicochemical configurations of each itemset. The latter can be achieved by grouping amino acids according to chemical properties, as shown later, and can help to identify isofunctional mutations or PSEs.

Algorithm 2: Pseudocode of the algorithm used for consensus calculation.

```
   Input: itemset observations 𝒲(I) of itemset I
   Output: consensus(I)
 1 begin
 2     initialize RMSD distance matrix D_{|𝒲(I)|×|𝒲(I)|}
 3     for i = 1 to |𝒲(I)| do
 4         D_{i,i} ← 0                                  // fill diagonal with zero
 5     end
 6     foreach (W_i, W_j) ∈ 𝒲(I) with i ≠ j do
 7         W_i' ← optimal superimposition of W_i onto W_j [136]
 8         D_{i,j} ← RMSD(W_i', W_j)                    // fill with pairwise distances
 9     end
10     consensus(I) ← 0
11     while |𝒲(I)| > 1 do
12         (W_i, W_j)_min ← min(D_{i,j} ∈ D, i ≠ j)
13         consensus(I) ← consensus(I) + D_{i,j}
14         W̄ ← average superimposed coordinates (W_i', W_j)
15         𝒲(I) ← 𝒲(I)\(W_i, W_j)_min                   // remove closest pair
16         𝒲(I) ← 𝒲(I) ∪ {W̄}                           // add consensus observation
17         recalculate D
18     end
19 end
```

Separation To solve the problem of obtaining itemset observations that are consecutive at the sequence level and thus exhibit a low cohesion or adherence, a further metric is introduced that enforces itemsets to be separated sequentially. The separation of residues in the protein sequence is another frequently observed property of structural motifs (see Chapter 3, Definition 3.1), which renders pure sequence-based analysis unfeasible. To mimic a repelling effect of items that are too proximal with respect to their sequence positions, a discretized adaption V_{Morse} of the Morse potential function [210] is used (Figure 6.10). This function is commonly used in physics to model the potential energy of diatomic molecules. To simulate repelling forces of items that are closely neighbored at the sequence level, an optimal desired gap length p_{opt} between items (minimum location of V_{Morse}) is defined:

$$V_{\text{Morse}}(p) = \alpha(1 - \exp(-\beta(p - p_{\text{opt}})))^2 - \alpha \tag{6.13}$$

A constant well depth of $\alpha = 500$ and a shape parameter of $\beta = 0.20$ were found to be suitable parameters to select itemsets close to p_{opt}. The score $S(W)$ of a single itemset observation $W \subset T = (t_1, ..., t_l, ..., t_m)$ can now be determined by mapping the position l of each item $w_i \in W$ to the ordered m-tuple T using the injective mapping function $P : W \mapsto T$ and calculating the discretized Morse potential V_{Morse} for successive items:

$$S(W) = \frac{1}{|W|} \sum_{i=1}^{i<|W|} V_{\text{Morse}}(P(w_{i+1}) - P(w_i)). \tag{6.14}$$

Subsequently, the separation of an itemset is defined as the average score of each itemset observation:

$$\text{separation}(l) = \frac{1}{|\mathcal{W}(l)|} \sum_{W \in \mathcal{W}(l)} S(W). \tag{6.15}$$

For itemset observations $W \in \mathcal{W}(l)$ with sequence gaps between items close to p_{opt}, the separation of l will be minimized with a theoretical global optimum of $-\alpha$ (Figure 6.10, Label 1). In contrast, itemset observations with items that fall below the optimal gap length p_{opt} at the sequence level are penalizing their associated itemset l resulting in an increased separation value (Figure 6.10, Label 2). If the gap length greatly exceeds p_{opt}, the score value of $S(W)$ will not be minimized further and $\lim_{p \to \infty} V_{\text{Morse}} = 0$ (Figure 6.10, Label 3). Hence, a neutral score is given to itemset observations with sequence separations considerably larger than p_{opt}. Negative separation values indicate itemsets with a sequence separation of at least p_{opt}. Separation should be minimized to discover itemsets not obvious at the sequence level due to non-contiguous items and long-range contacts of structural motifs.

MAPPING OF ITEM LABELS

To provide the possibility to consider similarities between amino acids during the mining process, a surjective mapping function $M : U \mapsto U'$ is introduced that redefines the universe of items and maps item labels to a new alphabet. Using the remapping of item labels prior to the mining process, it is possible to consider physicochemical similarities of amino acids, e.g. classification according to functional groups or chemical properties as proposed by [42] or [66]. The remapping of item labels is applied to each protein in the database:

$$\forall\, T \in \mathcal{T} : T \leftarrow M(T). \tag{6.16}$$

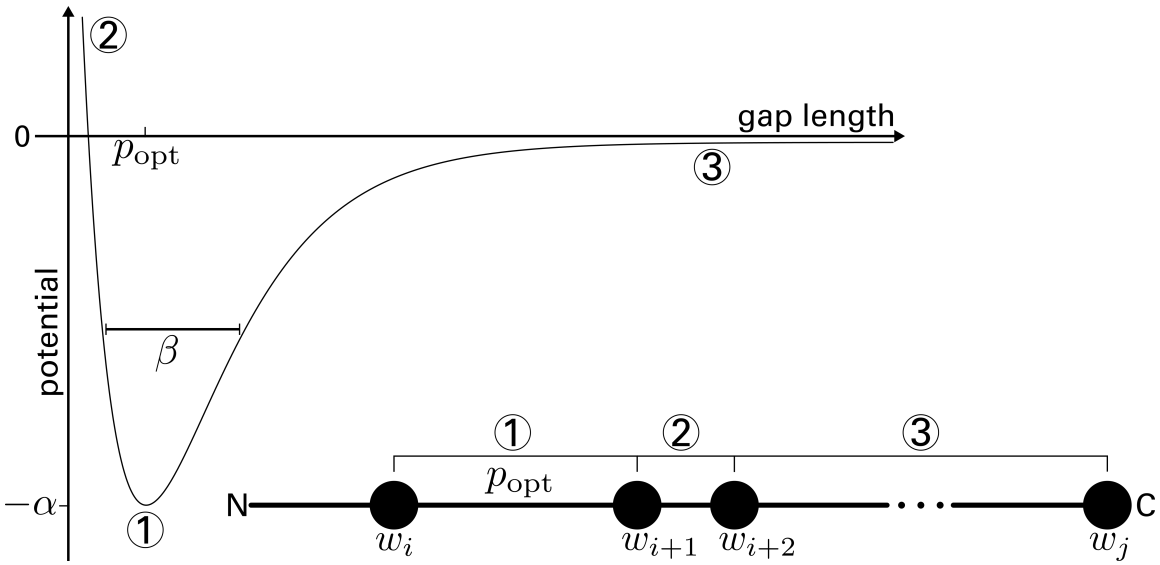


Figure 6.10.: Illustration of the separation metric and scoring of different gap sizes. The gap size between successive items (amino acids from N- to C-terminal direction) is evaluated with an adapted and discretized version of the Morse potential function [210] with its minimum $-\alpha$ at the desired ideal gap length p_{opt} and the shape parameter β. The separation value will be minimized if gaps close to p_{opt} are observed (**1**) and penalized if gaps less than p_{opt} are observed (**2**). Gaps larger than p_{opt} do not significantly contribute to a minimization of the separation value and a neutral score is assigned with increasing gap size (**3**).

Additionally, if the remapping reduces labels by grouping amino acids based on their properties, the number of possible itemsets is reduced and the mining process is accelerated.

STATISTICAL SIGNIFICANCE ESTIMATION

The estimation of the statistical significance of obtained itemsets is essential. To test whether the observed geometric similarity, determined by the consensus metric, occurs due to chance the statistical model proposed by MEYSMAN ET AL. was adapted [151]. After the mining process converged, all frequent itemsets $I \in \mathcal{I}$ are tested for significance. The bijective mapping function $N : U \mapsto \Pi$ maps the universe of items to a random permutation:

$$\Pi = \begin{pmatrix} u_1 & u_2 & \dots & u_i \\ \Pi(u_1) & \Pi(u_2) & \dots & \Pi(u_i) \end{pmatrix}. \tag{6.17}$$

Based on database entries with randomized labels, background distributions of the consensus value of each itemset I_Π are calculated by extracting itemset observations $\mathcal{W}(I_\Pi)$ from the randomized transactional data:

$$\mathcal{W}(I_\Pi) : \{W \in \binom{N(T)}{k} : L(W) = I_\Pi\} \; \forall \; T \in \mathcal{T}. \tag{6.18}$$

If this process is repeated several times, it is assumed that the obtained background distribution follows a normal distribution with a mean of μ and a standard deviation of σ:

$$\text{consensus}(I_\Pi) \sim \mathcal{N}(\mu, \sigma) \tag{6.19}$$

with

$$\mu = \frac{1}{n} \sum_{W \in \mathcal{W}(I_\Pi)} \text{consensus}(I_\Pi) \tag{6.20}$$

and

$$\sigma = \sqrt{\frac{1}{|\mathcal{W}(I_\Pi)|} \sum_{W \in \mathcal{W}(I_\Pi)} (\text{consensus}(I_\Pi) - \mu)^2}. \tag{6.21}$$

Subsequently, the quality of the modeled distribution is estimated using a Kolmogorov-Smirnov test [211] for the quality of continuous distributions. Only if the null hypothesis of this test is not rejected at significance level 0.1 the p-value for itemset I is calculated based on the cumulative probability of the background distribution.

ANNOTATION OF NONCOVALENT INTERACTIONS

The template-free detection of structural motifs largely benefits from the consideration of noncovalent interactions as shown for the HIGH motif in Class I aaRSs (Section 5.4.3). Interaction data annotated by PLIP [167] allowed to discover a conserved π-cation interaction between the two histidine residues of the motif. Thus, each database entry $T \in \mathcal{T}$ can be enriched with interaction data as follows. Consider int $= (a, b, u)$ to be a triple representing an interaction and consisting of a source coordinate $a \in \mathbb{R}^3$ and a target coordinate $b \in \mathbb{R}^3$. Further, an associated label for noncovalent interactions $u \in U_{\text{int}} = \{\text{hal}, \text{hyb}, \text{hyp}, \text{mec}, \text{pic}, \text{pis}, \text{sab}, \text{wab}\}$ is given. The labels[10] represent the interaction types supported by PLIP. The interactions are annotated using the established pseudocenter approach [144, 197, 212], where each interaction is added to its corresponding database entry T as new item $t = (u, v)$ with

$$v = \frac{a - b}{2}. \tag{6.22}$$

REFERENCE STRUCTURE COVERAGE

In order to assess the results of the mining process an intuitive visualization was designed that maps obtained results to a given reference structure that is part of the dataset. Consider $\hat{T} \in \mathcal{T}$ to be the selected reference structure. The coverage can be determined for each item (amino acid) $\hat{t} \in \hat{T}$ by calculating a coverage score as follows. For each itemset observation $W \in \mathcal{W}(I)$ of frequent and significant itemsets $I \in \mathcal{I}$ the structure of origin $O(W)$ is determined. Each item w of the observation W is mapped to its corresponding item $P(w)$ in the database entry of origin. The absolute coverage for each item $\hat{t}_i \in \hat{T}$ at position i of the reference structure can be calculated as follows:

$$\text{coverage}_{\text{abs}}(\hat{t}_i) = \sum_{I \in \mathcal{I}} \sum_{W \in \mathcal{W}(I)} \sum_{w \in W} \begin{cases} 1 & \text{if } O(W) = \hat{T} \text{ and } P(w) = \hat{t}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{6.23}$$

After calculating the coverage for each item of the reference structure, the values are normalized:

$$\text{coverage}_{\text{norm}}(\hat{t}) = \frac{\text{coverage}_{\text{abs}}}{\max\limits_{\hat{t} \in \hat{T}} \text{coverage}_{\text{abs}}(\hat{t})}. \tag{6.24}$$

This results in $\text{coverage}_{\text{norm}}(t') \in [0, 1]$ for each position of the reference structure describing how often this position is part of a frequent and significant itemset. This information is then stored in the B-factor [213] annotation of PDB files in order to exploit the

[10] **hal**: halogen bond, **hyb**: hydrogen bond, **hyp**: hydrophobic interaction, **mec**: metal complex, **pic**: π-cation, **pis**: π-stacking, **sab**: salt bridge, **wab**: water bridge

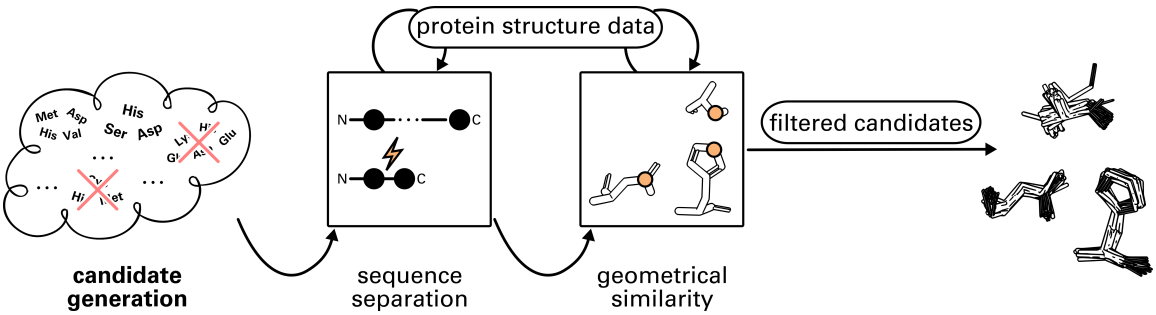


Figure 6.11.: Workflow of template-free structural motif detection with itemset mining and the extraction-dependent metrics separation and consensus. The Apriori algorithm [205] is used to generate candidates, which are then evaluated with biologically justified metrics. Simple evaluation metrics are not able to extract, nor to operate based on extracted itemset observations, but evaluate candidates against the database. Extraction metrics are able to extract concrete observations of itemsets from the database, which can be successively used by extraction-dependent metrics for further evaluation, e.g. regarding their level of geometric similarity (the consensus metric) or sequence separation (the separation metric).

visualization capabilities of the PyMOL software [147] for a per-residue depiction of the coverage score.

SUMMARY

An overview of the developed workflow for template-free structural motif detection is depicted in Figure 6.11. A three-step evaluation of itemset candidates is conducted according to:

1. evaluation metrics (support),
2. extraction metrics (cohesion or adherence), and
3. extraction-dependent metrics (consensus, separation).

The introduction of adherence, consensus, and separation allows the identification of structural motifs, which are separated in protein sequence, close to a desired spatial extent, and have a conserved geometry in a dataset of protein structures. Additionally, the grouping of amino acids according to physicochemical properties by remapping item labels can greatly reduce the search space for irrelevant itemsets. The method can be used to identify conserved structural motifs in an arbitrary dataset of protein structures.

6.2.3. BENCHMARK AND VALIDATION

The template-free structural motif detection with Fit3D was tested regarding its runtime and applicability. For this purpose, benchmark datasets and test cases were defined that highlight the ability of the method to detect structural motifs of functional relevance.

TIME COMPLEXITY

The algorithmic complexity of the proposed method depends on the size n of the input database, the number of individual labels $|U|$, and the metrics used for evaluation of candidates. In the following the usage of the metrics support, cohesion, separation, and consensus is assumed. The worst time complexity of the Apriori algorithm [205] is $\mathcal{O}(2^{|U|})$, i.e. all possible itemsets are generated and no evaluation thresholds for evaluation metrics are set. However, intelligent pruning of the search space and the use of biologically justified evaluation metrics adds an important constraint on this complexity. The support can be

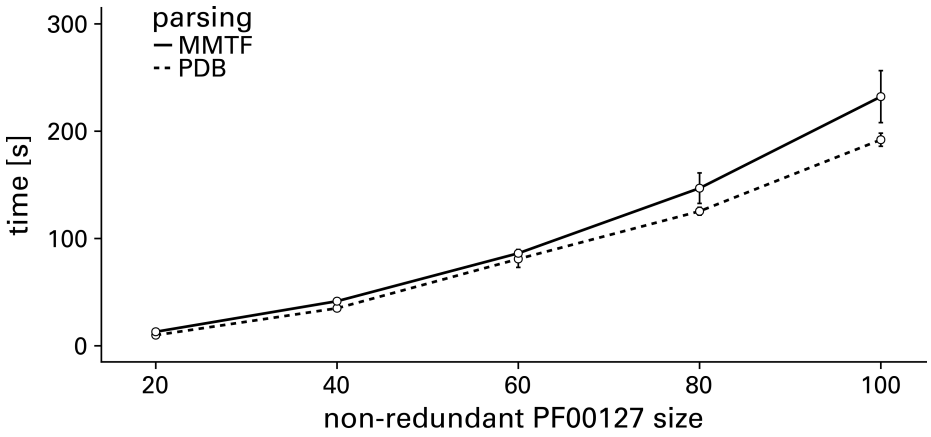


Figure 6.12.: Depiction of an averaged 5-fold runtime benchmark against datasets of non-redundant plastocyanin proteins of different size. The benchmark was carried out for parsing structures in PDB format or MMTF. The runtime includes a 5-fold significance estimation of the mined itemsets. Error bars indicate the standard deviation of five independent runs.

calculated within $\mathcal{O}(n)$, simply by counting the occurrence of each itemset in the target database of size n. The calculation of cohesion requires the determination of the smallest enclosing ball, which can be done in $\mathcal{O}(nm^2)$ for each itemset depending on the size m of each database entry. The heuristic VertexAll algorithm [150] is used to reduce the time complexity for the cohesion calculation. Evaluation of the separation metric takes $\mathcal{O}(n)$ time, simply be querying a hash table for the sequence positions of each itemset observation. Calculating the consensus score for each itemset takes maximal $\mathcal{O}(n^3)$ time [208], because at most n itemset observations are extracted for cohesion. Thus, the overall worst time complexity of the algorithm is exponential with $\mathcal{O}(2^{|U|} + nm^2 + n + n^3)$. There is evidence that mining maximal frequent itemsets is NP-hard [214]. However, as will be shown in the following section, the application of biologically justified evaluation metrics helps to reduce the search space and run the algorithm in reasonable time.

RUNTIME ANALYSIS

The performance of the method was evaluated by comparing the computation time for different sizes of non-redundant datasets and the parsing in PDB format or MMTF. The datasets consisted of proteins of the plastocyanin family of cupredoxins (Pfam:PF00127). The results were obtained and averaged for five independent runs for each dataset. The cohesion extraction metric was tested alongside with the extraction-dependent metrics consensus and separation. A comparison of the performance between PDB and MMTF regarding different dataset sizes is shown in Figure 6.12. Overall, the template-free detection of structural motifs for a dataset with 100 structures took approximately 192,064 ± 6,106 ms for PDB parsing and 231,221 ± 24,222 ms for MMTF parsing. Hence, the mining process takes approximately 1.9 s/structure if the PDB format is used and 2.3 s/structure if MMTF is used. Despite the faster parsing of structures with MMTF, a benefit for the overall runtime is not evident. The discrepancy between the two parsing strategies increases with larger dataset size. For 100 structures MMTF took on average 40,157 ms more time.

VALIDATION

The template-free detection was validated for two protein families that contain several structures in the PDB: trypsin serine proteases and cupredoxins. Serine proteases are well-studied enzymes that catalyze a peptide bond cleavage reaction by a charge-relay system consisting of histidine, aspartic acid, and serine [2]. The catalytic triad of serine proteases

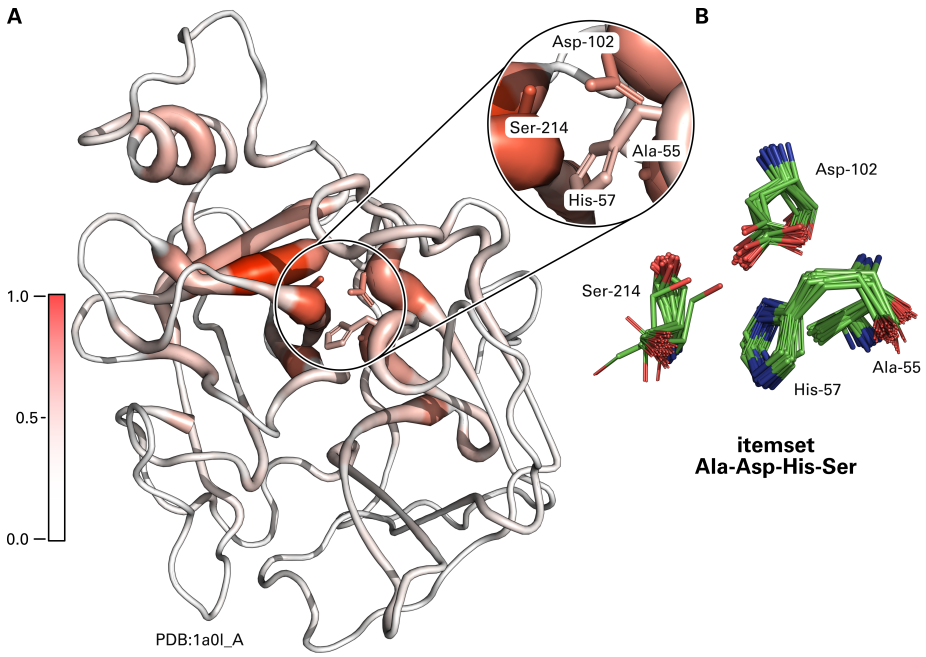


Figure 6.13.: Results of the template-free structural motif detection in the family of serine proteases (Pfam:PF00089). (**A**) A human tryptase enzyme (PDB:1a0l chain A) was used as reference structure. The color intensity and thickness of the loop representation corresponds to the coverage of conserved structural motifs at certain residue positions. The residues serine 214, aspartic acid 102, and histidine 54 constitute the catalytic site of the enzyme [215] and show a high coverage. (**B**) The corresponding itemset Ala-Asp-His-Ser, which contains the three catalytic site residues, shows a single dominant population of structurally conserved residues.

was already used for the validation of template-based detection in Section 6.1.3. Cupredoxins are copper-binding proteins and the plastocyanin family of cupredoxins is essential for the electron-transfer in photosynthesis and contains a type I copper binding site (T1Cu) composed of cysteine, methionine, and two histidines arranged in a distorted tetragonal geometry [107]. Structures associated with these families were derived from the Pfam database [132] (see Section 6.4). For both protein families the catalytic site residues could be extracted with high confidence.

Serine proteases Figure 6.13A shows the coverage of residue positions in the reference structure PDB:1a0l chain A, a human tryptase enzyme. The coverage score was obtained by mining a dataset of 116 non-redundant protein structures, derived from Pfam:PF00089. Red and bulky residues correspond to positions of high structural conservation. Serine at position 214 exhibits a high coverage score of 0.73, aspartic acid 102 of 0.50, and histidine 54 of 0.38, respectively. These three residues are part of the active site of the enzyme [215] and form the catalytic triad. Other residues with high structure coverage include isoleucine 227 (1.00), tyrosine 228 (0.85), isoleucine 103 (0.84), and alanine at position 55 (0.50). While alanine 55 and isoleucine 103 are in vicinity to the active site and thus likely to be structurally conserved, isoleucine 227 and tyrosine 228 belong to conserved sequence region in tryptase, trypsin and chymotrypsin enzymes [215]. The corresponding itemset Ala-Asp-His-Ser (significance p-value<0.001, Table A.3) is shown in Figure 6.13B. It features one population of structurally conserved residues. The ten top-scoring itemsets for the datasets are listed in Table A.3.

Cupredoxins As second validation of the method a dataset of structures belonging to the plastocyanin family (Pfam:PF00127) was analyzed. Figure 6.14A shows the coverage

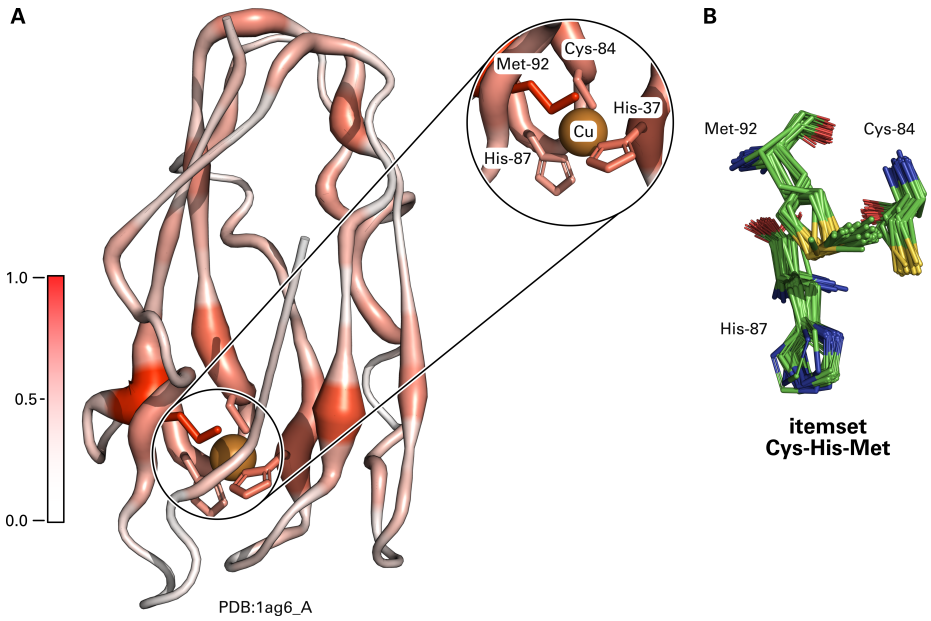


Figure 6.14.: Results of the template-free structural motif detection in the plastocyanin family (Pfam:PF00127). (**A**) A copper-bound plastocyanin of *Spinacia oleracea* was used as reference structure (PDB:1ag6 chain A). The color intensity and thickness of the loop representation corresponds to the coverage of conserved structural motifs at certain residue positions. Several residues of the copper-coordination center [216] consisting of histidine 37, histidine 87, cysteine 84, and methionine at position 92 were found to have a high structural conservation. (**B**) The corresponding itemset Cys-His-Met that contains three of the four ion-coordinating residues features a single population of structurally conserved residues.

of residue positions in the reference structure PDB:1ag6 chain A, a plastocyanin of *Spinacia oleracea*. The ten best scoring itemsets sorted by ascending consensus are shown in Table A.4. Residues methionine 92, histidine 37, cysteine 84, and histidine 87 exhibit a residue coverage of 1.00, 0.62, 0.52, and 0.45, respectively. All of these four residues constitute the T1Cu of the protein [107, 216]. Figure 6.14B shows the corresponding itemset Cys-His-Met (significance p-value<0.001, Table A.4) in its single dominant configuration of conserved geometry. Other residues with high structure coverage include lysine at position 30 and tyrosine at position 83. While lysine 30 was shown to be involved in crystal packing of the structure [216], tyrosine 83 may play an important role in the electron transfer between plastocyanin and cytochrome f; mutations of this residue lead to reduced activity [217].

6.2.4. IMPLEMENTATION

The detection of structural motifs based on itemset mining was implemented as command line tool as well as on API level. Currently, template-free detection is not part of the Fit3D web server. As for template-based detection, all implementations were written in Java and make use of the SiNGA [170] framework for the handling of macromolecular structures. Template-free detection supports a flexible computational representation of structural motifs as well as the definition of isofunctional exchanges via the usage of different categorizations of residues, e.g. by chemical groups [42].

Command Line Implementation The template-free structural motif detection algorithm was integrated in the Fit3D command line implementation[11]. The user can choose between the template-free and template-based mode of Fit3D. Due to a larger set of param-

[11] available at: github.com/fkaiserbio/fit3d/releases

eters for the template-free algorithm – compared to template-based detection – a predefined set is used when running from the command line. Alternatively, the user may define parameters freely by modifying a provided configuration file. To aid the interpretation of results, Fit3D requires to specify a structure that is used for coverage visualization. The default configuration of template-free detection includes the assessment of structural motif frequency (support), spatial vicinity (cohesion), sequence separation (separation), and geometric similarity (consensus). Instructions on how to use the implementation are provided in Appendix B.1.

API version Beside the easy-to-use command line version, a fully customizable implementation of the presented method is available at API level[12]. This version was designed in such a way that different evaluation metrics can be used in a "plug-and-play" manner (Figure B.2). It supports the specification of PDB identifiers or structure files as input. Furthermore, a single chain can be used as input and non-redundant structures are fetched automatically from the PDB REST API[13] according to a definable level of sequence similarity.

6.3. DISCUSSION

Adequate Computational Representation Reduces False Negatives Based on the results presented in Section 6.1.3 it was shown that an adequate computational representation of structural motifs plays an integral role when it comes to template-based matching in target structures. For the particular case of the ES, it was to be expected that multi-atom representation outperforms alpha carbon representation. This superfamily is known to have a high variance in alpha carbon positions and the backbone in general [33], which is why all-atom representation is key to increase specificity and sensitivity. The same could be observed for the NOS enzymes, where the template motif was defined based on the CSA according to [43]. In both cases all-atom representation outperforms alpha carbon representation with an increase in specificity of 25.84% for the ES and an increase in sensitivity of 51.88% for the NOS.

The results highlight how the consideration of full atomic resolution data reduces the risk to miss functional matches while simultaneously reducing the number of false positive matches. However, this statement cannot be generalized for every motif and has to be consolidated for every specific detection task. For instance, if variable regions of the motif are known or spatial variance is even favored, it may be reasonable to constrain atom representation to a custom selection of atoms. Although choosing the correct computational representation is crucial [195], alpha carbon representation was shown to be sufficient for some applications, e.g. to identify similar binding pockets [218]. There are other examples where such a single point representation does not suffice [33]. The missing feature to select arbitrary atoms for computational representation is a known deficiency of existing approaches and was addressed appropriately by Fit3D for both template-based and template-free structural motif detection.

Data Mining for Template-Free Structural Motif Detection The results presented in Section 6.2.3 show that the application of established data mining techniques is suitable for the template-free detection of previously unknown structural motifs in a set of protein structures. The catalytic triad of trypsin serine proteases as well as the copper ion coordination site of plastocyanin were detected without any *a priori* knowledge and by using

[12] available at: github.com/fkaiserbio/mmm
[13] see rcsb.org/pdb/software/rest.do, available as of May 18, 2018

biologically justified evaluation metrics for itemset mining. The introduction of adherence, consensus, and separation, contributes to a further optimization of the selection of candidates during the itemset mining process. Moreover, the consideration of isofunctional mutations or physicochemical similarities of residues by mapping labels to another alphabet allows covering structural motifs with limited sequence conservation. The representation of amino acids as chemical groups [42] showed to be suitable for the identification of conserved structural motifs in Class I aaRSs (see Section 5.4.3). Additionally, the execution time of the algorithm can be reduced if amino acids are grouped and the alphabet size is reduced. In conjunction with the consideration of noncovalent interactions, annotated by PLIP [167] and represented as pseusoatoms [197], subtle interaction properties of the HIGH motif in Class I aaRSs have been captured.

Introducing the separation metric allows to favor structural motifs with residues separated on the protein sequence level, as suggested by [151]. This optimization helps to avoid the detection of conserved geometries that arise due to secondary structure elements such as α-helices [219]. The separation score can be used as a good indication for long-range contacts, often observed for functionally important structural motifs which are hard to identify by sequence analysis. In conjunction with the consensus score this can lead to the identification of important residue patterns with a previously unknown function or to identify common elements of divergent protein families, e.g. superfamily templates [33]. Furthermore, the consensus approach can help spot different structural variants of motifs by allowing the hierarchical clustering of itemset observations. The usage of composite motifs, derived from averaging motif coordinates as used in the consensus approach, has been shown to improve protein function prediction [220]. Thus, the consensus representation of each cluster might be a valuable template motif to represent all cluster members. In general, the template-free detection algorithm considers all biological idiosyncrasies defined for structural motifs (see Chapter 3, Definition 3.1 and Definition 3.2):

- the spatial vicinity of residues via the cohesion concept,
- the separation at sequence level via the separation metric, and
- the geometric conservation via the consensus metric.

6.3.1. COMPUTATIONAL PERFORMANCE

Template-Based Detection The speed of template-based detection with Fit3D is substantially increased when using MMTF. This results from the parallel processing strategy of the implementation. Prior to a search, the structures of the target dataset are split into equal-sized subsets in respect to the number of available processor cores. Because the template-based detection can then be computed for each structure separately, parsing and computation can be run in parallel. Hence, fast structure parsing with MMTF accelerates the detection process to a great extent. Moreover, the number of expensive alignment operations is reduced by the selection of appropriate candidates as described in Section 6.1.2. The results of the runtime benchmarks indicate that template-based Fit3D is mainly applicable for small structural motifs with a size up to six residues, depending on the spatial extent of the template motif. This covers >90% of the structural motifs deposited in the CSA [124].

Template-Free Detection In contrast to template-based detection, the runtime of Fit3D in template-free mode did not experience a substantial benefit when using MMTF for structure parsing. This can be the result of several algorithmic peculiarities. Structure parsing is not performed in parallel and before the actual detection starts in order to be able to

perform all calculations on the whole dataset in memory. This is a constraint for the underlying itemset mining algorithm and the evaluation metrics, which have to be evaluated for the whole dataset in each round. Thus, caching all structure data is necessary. Due to the design of the MMTF standard, which is a compressed binary format [201, 221], the data structure is immutable. This raises another bottleneck when evaluating the consensus metric, which requires the calculation of a vast number of structural alignments [177]. In the current implementation of template-free Fit3D these alignments are represented by storing the rotation and translation information of coordinates alongside with the original coordinate data. This should be further optimized in future work such that recomputed coordinates of aligned structures are encoded binary and stored in MMTF.

6.3.2. BENEFITS OF FIT3D

As demonstrated in Section 5.4.2, template-based structural motif detection can be used to identify local similarities in protein structures for both globally similar and globally dissimilar properties. Especially the latter can be useful to, for example, repurpose known drugs where binding site similarity correlates with promiscuous drugs [222]. Hence, the assessment of binding site similarity is an important aspect of structural motif detection [168]. Many template-based methods, specifically designed to compare and identify defined ligand binding sites [44, 139, 218, 223], were developed. By exploiting template-based high-throughput detection, similarities of structural motifs can be revealed to predict and annotate protein function [121, 123], or to identify potential off-target binding sites during drug development [134, 224, 225]. Other template-based approaches [127, 140, 226] focus on the surface of proteins and the comparison or identification of patterns therein. Only a few generalizable methods exist that can be applied in order to screen for user-defined templates of structural motifs [129, 227, 228]. The Fit3D algorithm for template-based motif detection combines advantages of specialized methods and is applicable to be used for all mentioned problem categories. Although the focus of Fit3D lies on the detection of small structural motifs and the usage of full atom resolution, it should be a reasonable alternative to previous methods.

While many methods were developed to address template-based structural motif detection, the *de novo* identification of similarities in protein structures usually makes use of sequence information [154], or encompasses the abstraction of structural information [103, 105]. Methods relying on plain coordinate data of protein structures are rare. These methods were usually designed to be used with a small target dataset [157] or a pair of protein structures [158]. Here, the template-free detection engine of Fit3D clearly stands out and poses a novelty in the field. Firstly, Fit3D does not require any additional assumptions of the use case, such as ligand binding site detection or surface patch recognition. Secondly, by relying on the original coordinate data, Fit3D is independent of any sequence alignments or data transformation strategies. This results in the template-free identification of geometrically similar structural motifs that can be located anywhere in the protein structure and are not necessarily responsible for the interaction with ligands or catalysis. This could further contribute to the identification and understanding of structure-stabilizing elements [30] and PDB-wide molecular building blocks [151]. The presence of these hydrophobic motifs suggests that they are involved in hydrophobic core formation during protein folding [229–231]. Because Fit3D was designed with general-purpose applications in mind, further investigation of such patterns and the analysis of their geometric characteristics is reasonable and could be fostered by Fit3D.

Nevertheless, the identification of regions that are important for a specific protein's function, as demonstrated for trypsins and cupredoxins, is promising. To the best knowledge of the author, Fit3D is the first application of an adaption of itemset mining on protein

structures that utilizes geometric similarity and sequence separation to identify common structural motifs in arbitrary-sized sets of input data. The results demonstrate how established data mining techniques can help to shed light on biological data. However, the choice of an appropriate method ultimately depends on the case of application, e.g. the desired computation time and accuracy, or the focus of the study, e.g. on ligand binding sites at the protein surface.

Fit3D: A General-Purpose Tool Even if the analysis of protein structures is the focus of this thesis, applications of the developed algorithms on other types of macromolecular structure data are conceivable. These might encompass, for example, the template-based detection of structural motifs in riboswitches [160] or the mining of trajectories of molecular dynamics simulations for especially stable regions using template-free detection. Fit3D's template-free mode might also help to study the flexibility of protein structures, e.g. when applied on nuclear magnetic resonance models or the results of normal mode analysis [232].

The usage of state-of-the-art technologies ensures that Fit3D can keep pace with the rapidly increasing number of available protein structures in the PDB. Improvements in experimental structure determination methods will lead to high-resolution data for large protein structures and macromolecular complexes. Both aspects are already covered by Fit3D; the support of future-proof data formats such as MMTF and the consideration of all-atom data via the flexible computational representation of structural motifs. The template-based structural motif search with Fit3D is almost parameter free, which is a strong plus of the method. Even though template-free structural motif detection requires a decent set of parameters, the parameters used for the analysis in this thesis have proofed to be a suitable choice. The availability of an API version of Fit3D provides the possibility to integrate the method into sophisticated processing pipelines, which meet the requirements of expert users. The flexibility of Fit3D was demonstrated for the specific analysis of aaRS structures.

Features of Fit3D Fit3D combines the advantageous features of already existing methods into a single tool for both template-based and template-free structural motif detection (see Table 6.1). Template-based detection uses a combinatorial algorithm, while template-free detection relies on itemset mining to detect structural motifs. No general limitations for a specific usage scenario, such as binding site comparison or surface patch recognition, exist. However, both technologies have some minor limitations. The combinatorial algorithm can be time-consuming and works best for small structural motifs up to a limited size and spatial extent. Motifs with a large spatial extent lead to the extraction of large local environments (see Algorithm 1, Line 5) which in turn lead to many match candidates that have to be tested for similarity by superimposition. If larger motifs should be processed, a splitting into subsets of residues is advisable before applying Fit3D. Because structural motif detection is closely related to the subgraph isomorphism problem [46, 47], it can be assumed that no exact polynomial time algorithm exists to solve the problem. A main limitation of the template-free detection originates from the definition of itemsets (Definition 6.1) where the repetition of labels (amino acids) is prohibited according to the set definition. Only through this algorithmic constraint in itemset mining it is possible to explore the candidate search space in reasonable time. This can result in difficulties of the method to discover structural motifs with repetitive amino acids, often the case for metal ion coordination centers such as the popular zinc finger motif consisting of two cysteine and two histidine residues [119]. However, as shown for cupredoxins (Section 6.1.3) where two histidine residues are involved in ion binding [107], the structural motif could still be identified with high confidence. The algorithm does not guarantee which histidine

Table 6.1.: An overview of the features available with Fit3D. Template-based detection follows a combinatorial (CO) approach, while template-free detection relies on itemset mining (IM). Features marked with an asterisk are only available in the API version of the implementation.

			features						use		
		reference	custom atom representation	inter-molecular motifs	PSEs	statistical significance	DNA/RNA motifs	ligand motifs	implementation	open source	limitation
CO	Fit3D template-based	[195]	✓	✓	✓	✓	✓	✓	✓	✓	small structural motifs
IM	Fit3D template-free	[177]	✓	-	-	✓	✓	✓*	✓	✓	no amino acid repetitions

residue will be used for the comparison because the observation with the smallest enclosing ball radius per structure will be selected. This nondeterministic behavior results in the detection of both motif residues. Hence, the coverage score proposed in Section 6.2 allows for a visual and quantitative identification of structural motifs beyond the limitation of itemsets.

6.3.3. LIMITATIONS OF GEOMETRIC APPROACHES

Ligand Binding Site Similarity Fit3D is a geometric detection method that relies on the superimposition of three-dimensional coordinates. In the context of this thesis, and the definition of structural motif conservation (Chapter 3, Definition 3.2), this is an integral requirement for structural motifs. However, there are scenarios where assessing geometric similarity is insufficient, e.g. for the identification of alternative drug targets. The recognition of an identical ligand in proteins can be based on several mechanisms [190]:

- similar proteins recognizing the ligand with similar binding sites (Figure 6.15A),
- different proteins recognizing the ligand with similar binding sites (Figure 6.15B),
- similar proteins recognizing the ligand with different binding sites (Figure 6.15C), and
- different proteins recognizing the ligand with different binding sites (Figure 6.15D).

Hence, the assessment of geometric similarities can be challenging [4, 168] and may miss potential target predictions if the binding sites are dissimilar at the geometric level (Figure 6.15C and Figure 6.15D). There are several examples where an identical ligand is recognized via different mechanisms. For example, adenosine phosphate ligands are known to be bound by the P-loop domain [190, 233], positive charges [42], or backbone hydrogen bonding [163, 234]. Other examples include aromatic ring detection via π-stacking [190] or π-cation interactions [184]. According to the "functionalist principle" in biology [4], the structure of proteins and binding sites can differ and evolve divergently as long as the function is not compromised.

Yet, the geometric conservation of structural motifs can be strong as shown for Class I and Class II aaRSs, where evolution conserved identical ATP binding geometries despite a strong divergence in global sequence and structure [163]. This shows that geometrically similar structural motifs can be even found in subsets of dissimilar binding sites. In general, there seems to be no single code for the identification of similar binding sites in unrelated

proteins [190, 235], but there might be a limited number of different patterns to recognize functional groups of the ligand [42, 236]. It can be reasonable to consider noncovalent interaction [168, 237] data for the definition of "generic structural motifs" independent of the concrete amino acid implementation and residue geometry. This was addressed in this thesis by including intrinsic protein interaction data, mapped to pseudoatoms, during the template-free structural motif detection in Class I aaRSs (Section 5.4.3). However, as Fit3D relies on the superimposition of coordinates only geometrically conserved interactions can be identified.

Conformational Changes Upon Ligand Binding In addition to different binding mechanisms, the dynamics upon ligand binding are still not fully understood. Theories explaining the binding process include the rigid key-lock principle, the induced fit mechanism [238] (debated by [239]), or the conformational selection and population shift hypothesis [240, 241]. Despite the actual mechanisms are still a matter under discussion and seem to depend on the individual example, conformational shifts were shown to have a crucial role on enzyme function [242, 243]. During evolution, enzymes have likely developed from conducting large, energy-consuming motions such as backbone rearrangement, towards side-chain flexibility to facilitate catalysis and consume as little energy as possible [186, 244]. Hereby intrinsic binding site interactions play a crucial role; polar residue-residue interactions in the binding site are an indicator for more rigid systems, whereas non-directional aromatic and hydrophobic interactions tend to occur in flexible binding sites [186]. In terms of the movement of individual residue side chains in the binding site, polar amino acids tend to be more flexible compared to aromatic residues [185]. All these dynamic factors have to be kept in mind when performing structural motif detection. Nevertheless, due to the flexible computational representation of structural motifs in Fit3D, different conformations of structural motifs can be identified. If, for example, high side chain variance is observed between ligand-bound and ligand-free state as for the Arginine Tweezers motif (Section 5.4.1), the representation of the structural motif via backbone atoms is feasible. If, on the other hand, displacements of the protein backbone are observed, e.g. for the ES template motif [33], it is reasonable to consider all atoms for structural motif detection in order to compensate these movements.

6.4. MATERIALS AND METHODS

Template-Based Detection Runtime Benchmark The datasets used to test the runtime of template-based structural motif detection were created as follows. The serine protease template motif was derived from the CSA [124] based on the structure PDB:1gl0 chain A, and consisted of residues histidine 57, aspartic acid 102, and serine at position 195, respectively. The target structure datasets were composed by random selection of structures from a non-redundant PDB snapshot as of April 16, 2016 and a BLAST *p*-value of 10^{-7} according to VAST [245]. Subsequently, all structures were checked to be available in PDB and MMTF format. The benchmark was executed on a standard Intel™ Core i7-6700 CPU machine equipped with 32 GB of RAM and an SSD drive. Each benchmark case consisted of five warm-up iterations as well as five measurement iterations and was implemented with the Java Microbenchmark Harness (JMH) framework.

Template-Based Detection Validation The datasets used for the validation of template-based detection with Fit3D were derived from the CSA [124] and the SFLD [202]. The template motif was derived from the primary CSA entry PDB:3nos chain A of the NOS

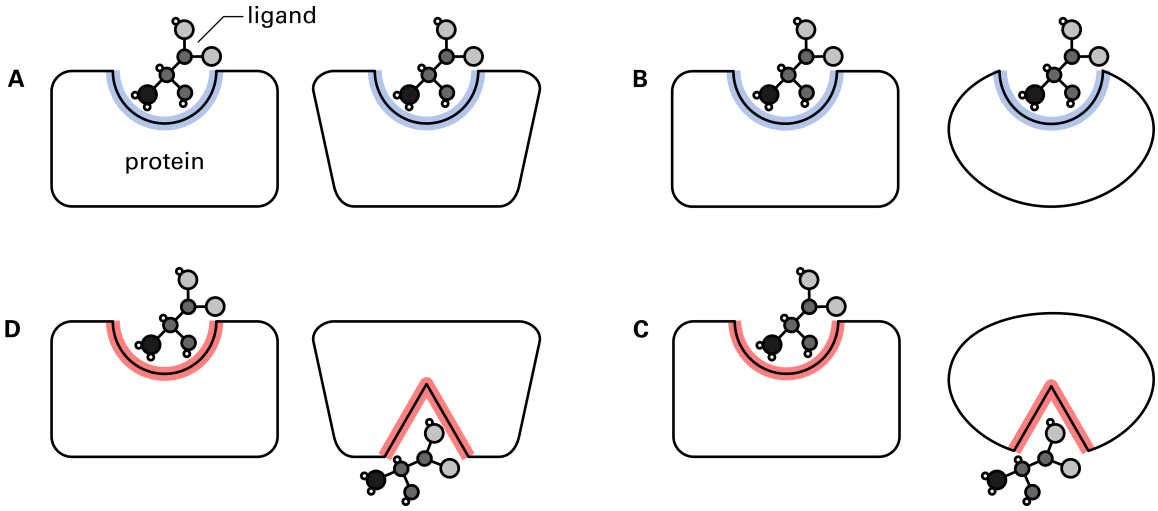


Figure 6.15.: The recognition of an identical ligand. (**A**) Similar proteins that recognize the ligand with similar binding sites and at the same ligand moiety. (**B**) Different proteins that recognize the ligand with similar binding sites and at the same ligand moiety. (**C**) Similar proteins that recognize the ligand with different binding sites at a different ligand moiety. (**D**) Different proteins that recognize the ligand with different binding sites at a different ligand moiety. Structural motif detection algorithms are only able to identify geometrically similar binding sites (**A** and **B**, emphasized in blue) and fail to identify dissimilar binding sites (**C** and **D**, emphasized in red) still binding at a different ligand moiety or in a different mode. Figure adapted from [190].

family (EC:1.14.13.39) and consisted of the residues cysteine 184, arginine 187, tryptophan 356, and glutamic acid at position 361. The positive dataset consisted of all 266 individual chains annotated in the CSA for this enzyme family. The positive dataset featured an average sequence identity of 70.44% after a global alignment [165] and an average TM score of 0.94 [164] (Figure A.1). It was not filtered for redundancy to be consistent with the dataset presented in [43]. Only exact matches of the residues specified to be catalytically active were counted as true positive at significance level p-value<0.001 [161]. The negative dataset contained 39,562 protein chains and was derived from a non-redundant PDB snapshot as of April 16, 2016 and a BLAST p-value of 10^{-80} according to VAST [245]. All structures present in the positive dataset were excluded from the negative dataset. The template motif for the detection in ES was extracted from PDB:2mnr chain A according to [33] and PSEs were defined as follows: lysine 164 exchangeable with histidine, glutamic acid 247 with aspartic acid and asparagine, and histidine 297 with lysine. The positive dataset was composed of non-redundant protein chains of the ES (SFLD:1) as defined in the SFLD and filtered for redundancy with a p-value of 10^{-80} according to VAST [245]. The resulting subset of structures showed an average pairwise sequence identity of 23.17% [165] and an average TM score of 0.81 [164] (Figure A.2). The ES structural motif as defined in [33] was mapped to this extended set of structures by sequence alignment [165] of all structures against structures in the MENG ET AL. dataset. This resulted in 73 protein chains with the structural motif mapped to sequence positions instead of the original 43. The negative dataset was identical to that used for NOS and all structures of the positive dataset were excluded.

Template-Free Detection Runtime Benchmark To test the runtime of template-free structural motif detection with Fit3D, the algorithm was run for datasets of the plastocyanin family (Pfam:PF00127) of different size. These datasets were created by randomly selecting structures from the full dataset containing 105 entries. The parameters used for the runtime benchmark were identical to those used for the validation of the method (listed in the next paragraph). The benchmark was executed on a standard Intel™ Core

i7-6700 CPU machine equipped with 32 GB of RAM and an SSD drive. Each benchmark case consisted of five warm-up iterations as well as five measurement iterations and was implemented with the JMH framework.

Template-Free Detection Validation For the validation of template-free structural motif detection each dataset was filtered such that only single-chain proteins were considered and a non-redundant version with a BLAST *p*-value of 10^{-80} from VAST [245] was derived. Structures were selected from the Pfam database version 31 [124] with the identifiers Pfam:PF00089 and Pfam:PF000127 for trypsin and plastocyanin, respectively. The trypsin dataset contained 116 structures, whereas the plastocyanin dataset contained 105 structures. The parameters for both runs were set as follows: maximal support of 0.90, maximal cohesion of 5.00 Å, maximal separation of 100.00 and optimal separation of 5.00, maximal consensus of 0.60 with λ=0.50. Amino acids were represented in the three-dimensional space by all atoms excluding hydrogen. Significance estimation was based on 5-fold shuffling of item labels.

Implementation All algorithms were implemented with Java version 1.8 and utilize the SiNGA framework version 0.3.3 [170]. Beside structure parsing in PDB format, the parsing with MMTF [201] version 1.0 is supported. The Fit3D web server for template-based detection was implemented with JavaServer Faces version 2.2 and PrimeFaces version 6.1 and is running on a Tomcat 8 application server. For technical details on the implementations please refer to Appendix B.

7. CONCLUSIONS

Structural Motifs Might Have Shaped the Genetic Code The region reconstructed by MARTINEZ-RODRIGUEZ ET AL. [19], called Protozyme, was the minimal functional aaRS unit required in ancient protein biosynthesis. This region contains the N-terminal residue of the Backbone Brackets and Arginine Tweezers motif [163] (Figure 7.1). This suggests that both N-terminal residues can fulfill their functional role in isolation, but with reduced efficiency. During evolution, the aminoacylation reaction was further improved by adding their other functionally equivalent counterpart. According to the Rodin-Ohno hypothesis [18] (Chapter 2, Section 2.4), one can conclude the following chronological appearance of the Backbone Brackets and Arginine Tweezers motif. The N-terminal residues of both motifs seem to be the most ancient parts, both located in the Protozyme region. Over a prolonged period the C-terminal Backbone Brackets residue, which is located close to the KMSKS motif and hence part of the Urzyme, was introduced. The most recent residue seems to be the C-terminal Arginine Tweezers residue, located in Motif "3", which is neither part of the Protozyme nor the Urzyme. If the peptide-RNA world hypothesis [15, 18–20] holds true, it is conceivable that the two structural motifs were directly involved in the formation of the genetic code. By ensuring consistent ligand interaction, both structural motifs were major determinants for the evolvability of aaRSs.

Structural Motif Analysis Uncovers Subtle Aspects Furthermore, the strong structural conservation of both motifs in aaRS structures across all kingdoms of life and the ability to detect these structural motifs even in contemporary aaRSs, is an outstanding example of the evolutionary conservation of structural motifs. There might be other examples in the proteome, where fundamental structural motifs exists that are yet to be identified. The methods provided in this thesis can accelerate further analysis, especially in conjunction with the release of new experimental structure data. Conclusively, Open Problem I (defined in Chapter 1) was successfully addressed by applying new structural motif detection algorithms:

Results for Open Problem I

The structural motif detection algorithms developed in this thesis were used to identify and thoroughly characterize structural motifs in Class I and Class II aaRSs at atomic level.

The high-throughput template-based detection of the Arginine Tweezers and Backbone Brackets motifs in the PDB identified similar structural motifs, which might constitute an evolutionarily independent solution to the same biological problem of ligand fixation. The findings are strengthening the assumption that there is only a limited number of different patterns to recognize functional groups of ligands [42, 236]. For the analysis of aaRS structures, the characterization of the two structural motifs with Fit3D was shown to be sufficiently sensitive to suggest the structural rearrangement of Class I aaRSs to be a general mechanism. Hence, if structural motifs conserved in a larger number of protein structures are known, structural motif analysis can reveal insights into global structural effects that occur during ligand binding without requiring any additional information.

Fit3D: Enhanced Structural Motif Detection In the context of this thesis, enhanced algorithms for template-based and template-free structural motif detection were developed and validated. The results show that the adequate computational representation of structural motifs – not considered by most existing methods – is essential to achieve high sensitivity and specificity. Furthermore, algorithmic robustness of the template-based detection with Fit3D is guaranteed due to the small number of required parameters. Despite the fact that computation time is increased when considering geometric similarity at the

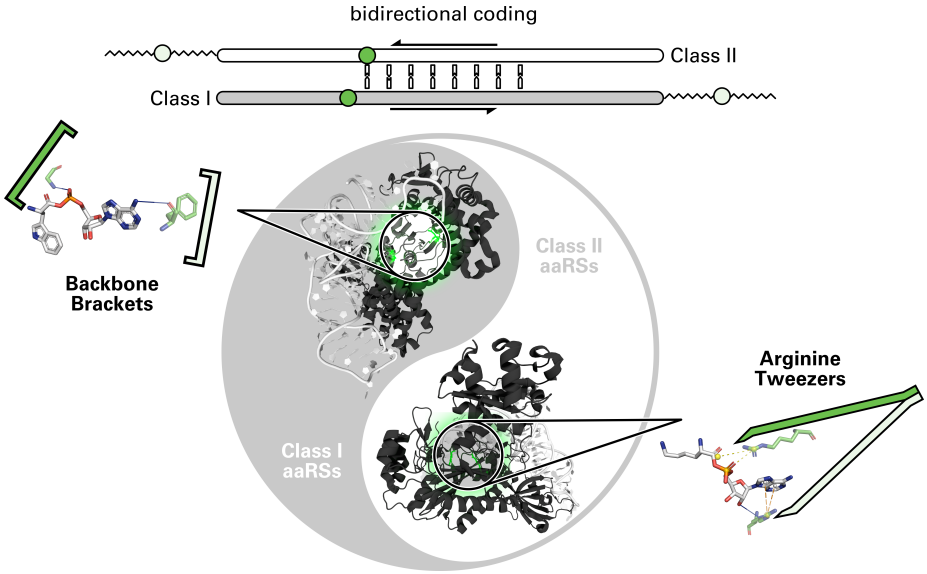


Figure 7.1.: The two distinct classes of aaRSs constitute a self-referencing system and feature oppositional implementations of ligand binding. In this thesis, two structural motifs were characterized with Fit3D: the Backbone Brackets and the Arginine Tweezers. Structural analysis revealed a strong conformational shift upon ligand binding for Class I aaRSs. The results show a stunning structural conservation of these motifs and thus suggest that they are present since ancient times. According to the Rodin-Ohno hypothesis [18] the primordial forms of aaRSs were presumably encoded bidirectionally on opposite strands of the same gene from which the contemporary classes originated. Both structural motifs were traced back to these primordial forms [163], indicating that they had an important role for the formation of the genetic code.

atom level, structural motif detection with Fit3D can still be conducted on datasets containing thousands of protein structures in reasonable time.

The novel algorithm was implemented in an easy-to-use command line software tool, which was released under the terms of public licensing. The tool is freely available and ready to be optimized, adapted, or expanded by the scientific community. Fit3D complements existing web services and standalone tools (e.g. [128, 130, 143, 246]). Furthermore, to the best knowledge of the author, it is the first structural motif detection software for which a fully-fledged API is available. This allows users to employ the software for very specific applications, as demonstrated for the analysis of structural motifs in aaRSs (Chapter 5). Thereby its application is a double-sided approach: on the one hand it is possible to search for structural motif templates in a large set of target proteins, e.g. in form of a CSA-derived library like done by [31]. On the other hand one can screen datasets of structures without the need for a specific template to discover new functionally important structural motifs. By extending existing concepts [150, 151], a new template-free detection algorithm was designed that uses actual geometric information, which is difficult to realize with previous methods. Libraries of structural motifs identified by template-free detection with Fit3D, could be applied to classify protein family associations [153] independent of sequence alignments and based on local structure. Furthermore, the usage of such libraries is conceivable to predict protein function [102].

Due to the rapid growth of automated structure determination methods through structural genomics effort, protein structures are often solved prior to biochemical and functional characterization [247]. Hence, Fit3D can proof to be very useful for scientists dealing with protein crystallography and function determination, which is especially important if novel folds are observed [40]. The field of drug design and research is also addressed by this approach: the mechanism of drug effect often lies in the inhibition of protein active sites, which are in turn describable through structural motifs. According to the "functionalist principle" it can be stated that three-dimensional information is more conserved than se-

quence and the determinant for physicochemical properties in binding sites, which is why structural motifs are conserved [4, 123]. Additionally, ligand binding site similarity was observed even if global sequence or structure similarity cannot be detected [225]. It is of high interest to investigate active site mechanisms and structures in a high-throughput manner. The application of structural motif detection algorithms can aid researchers to discover off-target binding of specific drugs [224, 225]. The Fit3D implementation can be seen as a valuable tool in the field of structural bioinformatics research and computational biology.

Structural Motif Detection Algorithms are a General Tool The algorithms developed for template-based and template-free structural motif detection are not limited to protein data. Fit3D can be applied to other types of macromolecular structure data such as DNA or RNA. The integration of noncovalent interaction data [167] may allow to address some limitations of geometric approaches and is one step towards the definition of "generic structural motifs". Thus, Open Problem II (defined in Chapter 1) was addressed by the development of new structural motif detection algorithms:

Results for Open Problem II

Two general-purpose algorithms for template-based and template-free structural motif detection were developed and validated. Open source implementations are provided to the scientific community as command line version, web sever, and flexible API.

In principle, the developed algorithms are not limited to biological data but can be applied to any type of labeled spatial data. This constitutes a contribution to computer science beyond the field of structural bioinformatics. For example, the approaches presented in this thesis could be relevant for computer vision to align three-dimensional point clouds [248] or to discover patterns therein [249]. Due to its open source code base, Fit3D is ready to be adapted and applied by the scientific community.

Structural Bioinformatics of Tomorrow The integration of state-of-the-art technologies, such as MMTF [201] as future-proof data format for macromolecular structures in the PDB, ensures that the Fit3D processing pipeline can keep pace with the steady increase of available structure data. The development of MMTF is funded by the Big Data to Knowledge (BD2K) initiative of the National Institutes of Health to "*[...] maximize and accelerate the integration of big data and data science into biomedical research.*"[14]. The BD2K program encompasses $200M of funding to accelerate diverse technologies for biomedical data science. Hence, the demand for novel algorithms in structural bioinformatics will experience a strong increase over the next years. The contribution of Fit3D is a small piece of the puzzle to support the analysis of tomorrow's biological data.

[14] commonfund.nih.gov/bd2k, available as of May 17, 2018

Part III.

APPENDIX

A. SUPPORTING INFORMATION

Table A.1.: The ten best scoring itemsets for aaRS Class I structures sorted by ascending consensus score. Residues were categorized according their chemical groups [42]. The separation score cannot be calculated for itemsets that exclusively consist of interactions.

itemset	significance	KS value	support	cohesion	consensus	separation
hyb-pic-imi-oth	***	0.9413	0.9136	3.5026	0.2381	42.7929
hyb-pis-imi-oth	***	0.6684	1.000	3.3568	0.2637	7.9766
hyb-pic-pis-oth	***	0.9719	0.9136	4.7052	0.3022	n/a
hyb-pic-pis-imi-oth	***	0.9558	0.9136	5.3940	0.3349	24.4131
hyb-pis-sab-imi	***	0.6264	1.000	4.0882	0.3559	n/a
hyb-pic-pis-car-oth	***	0.8480	0.9136	5.5241	0.3874	-4.1649
hyb-pis-sab-hyd-imi	***	0.8450	1.000	4.6906	0.3879	-16.9619
hyb-pis-sab-amd-imi-oth	***	0.9598	1.000	5.3797	0.3879	-13.6703
hyb-pic-pis-imi	***	0.9005	0.9136	5.2959	0.3972	n/a
hyb-pic-pis-sab-car-imi-oth	***	0.8052	0.9136	6.6121	0.4083	-1.6985

*** p-value<0.001

Table A.2.: The ten best scoring itemsets for aaRS Class II structures sorted by ascending consensus score. Residues were categorized according their chemical groups [42].

itemset	significance	KS value	support	cohesion	consensus	separation
hyb-pis-imi-oth	***	0.7508	0.9737	3.5253	0.3565	-11.3586
hyb-pis-sab-car-imi-oth	***	0.9573	0.9737	5.0544	0.3820	-4.8712
hyb-pis-sab-car-hyd-imi-oth	***	0.8899	0.9737	5.4197	0.3927	7.2263
hyb-pis-sab-imi-oth	***	0.6224	0.9737	4.7709	0.3980	-8.3195
hyb-pis-sab-car-imi	***	0.9772	0.9737	4.9503	0.3998	-28.5699
hyb-pis-sab-car-hyd-imi	***	0.2657	0.9737	5.3559	0.4027	-30.9885
hyb-pis-sab-car-gua-imi-oth	***	0.9959	0.9737	5.6096	0.4068	-16.0354
hyb-pis-sab-amd-car-hyd-imi-oth	***	0.8408	0.9737	5.8124	0.4098	22.3842
hyb-pis-sab-hyd-imi-oth	***	0.2783	0.9737	5.1614	0.4155	3.9023
hyb-pis-sab-amd-car-imi-oth	***	0.6296	0.9737	5.5958	0.4201	9.2515

*** p-value<0.001

Table A.3.: The ten top-scoring itemsets found in serine proteases (Pfam:PF00089) sorted by ascending consensus score.

itemset	significance	KS value	support	cohesion	consensus	separation
Ala-Asp-His	***	0.9516	0.9397	3.6790	0.2524	-43.1563
Gly-Leu-Pro	***	0.7989	1.0000	3.2111	0.2834	4.3016
Ala-Asp-His-Ser	***	0.9999	0.9397	4.4899	0.3014	-40.7461
Gly-Ile-Pro	***	0.6176	1.0000	3.2723	0.3037	20.7713
Ala-Asp-Ile	***	0.7984	0.9914	3.3673	0.3317	30.8890
Ala-Cys-Pro	***	0.9982	0.9569	3.7900	0.3343	18.9567
Ala-Pro-Val	***	0.8472	1.0000	3.2788	0.3352	4.9453
Asp-Ile-Trp	***	0.8440	0.9483	3.9206	0.3483	16.8371
Ala-Asp-His-Thr	***	0.3577	0.9397	4.8021	0.3493	-1.3426
Ala-Gly-Leu-Pro	***	0.5986	1.0000	3.9830	0.3503	19.2821

*** p-value<0.001

Table A.4.: The ten top-scoring itemsets found in plastocyanin proteins (Pfam:PF00127) sorted by ascending consensus score.

itemset	significance	KS value	support	cohesion	consensus	separation
Cys-Gly-His	***	0.8140	1.0000	4.4030	0.1510	-99.0408
Asn-Cys-Gly-His	***	0.9609	1.0000	4.8145	0.1557	54.7009
Cys-His-Met	***	0.9526	1.0000	3.2795	0.1667	-254.4509
Asn-Cys-Thr	***	0.9905	1.0000	4.7965	0.1716	21.8315
Cys-His-Pro	***	0.9925	1.0000	3.9309	0.1717	-74.5600
Cys-Gly-His-Pro	***	0.7773	1.0000	4.8509	0.1724	-3.5290
Asn-Cys-Tyr	***	0.8357	0.9905	4.5960	0.1748	-3.3693
Asn-Cys-Tyr-Val	***	0.5750	0.9905	4.9227	0.1804	-9.8418
Asn-Cys-Pro	***	0.4171	1.0000	4.9665	0.1808	-77.4633
Cys-His-Val	***	0.5845	1.0000	4.1965	0.1830	-69.3767

*** p-value<0.001

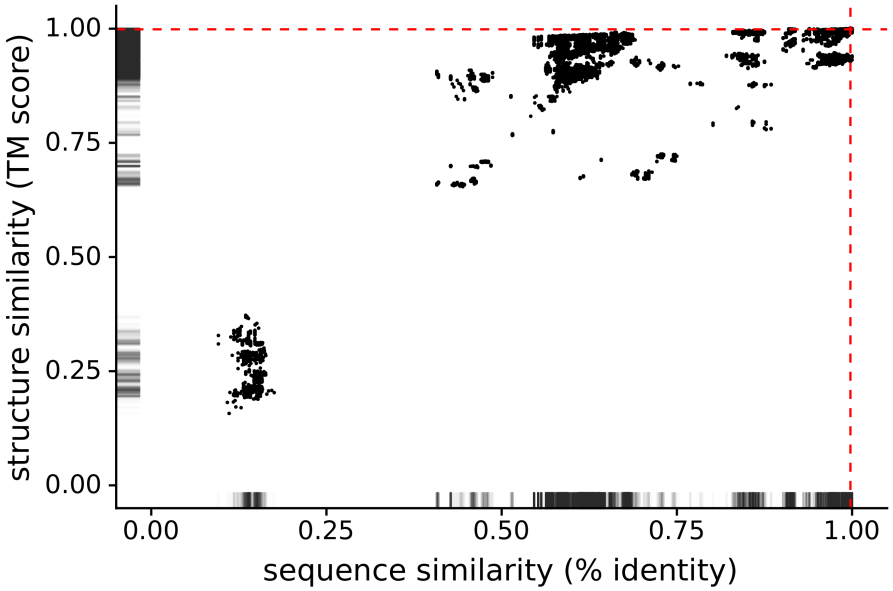


Figure A.1.: Pairwise sequence similarity (% identity after global sequence alignment [165]) versus pairwise structure similarity (TM score, [164]) of the NOS dataset. The 95%-percentiles are depicted by red dashed lines. This dataset was used as positive control for the experiments performed in Section 6.1.3.

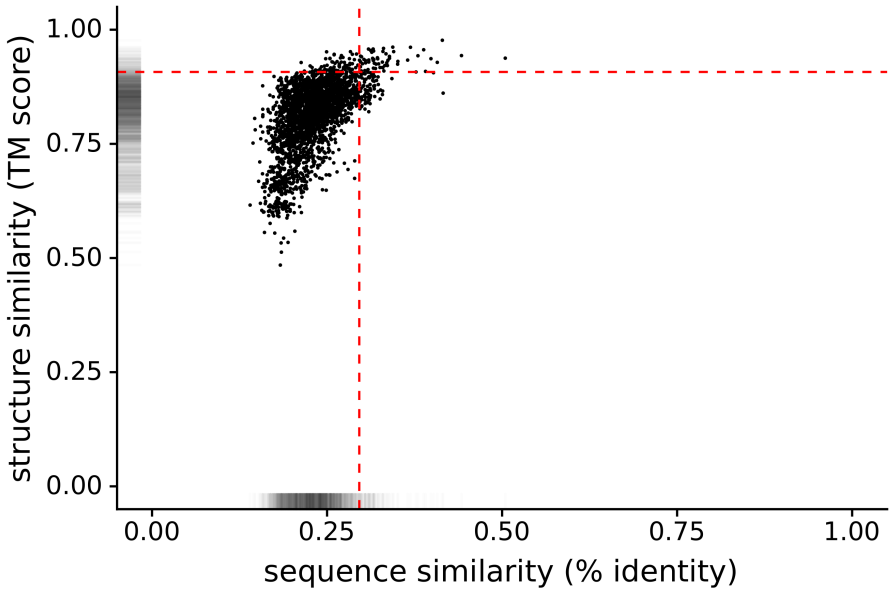


Figure A.2.: Pairwise sequence similarity (% identity after global sequence alignment [165]) versus pairwise structure similarity (TM score, [164]) of the ES dataset. The 95%-percentiles are depicted by red dashed lines. This dataset was used as positive control for the experiments performed in Section 6.1.3.

B. FIT3D TECHNICAL DOCUMENTATION

B.1. COMMAND LINE VERSION

The command line version of Fit3D was implemented in Java version 1.8 and uses the SiNGA framework [170] as well as the MMTF API [201]. This eliminates the need for other software dependencies because SiNGA features a comprehensive handling of macromolecular structure data. Furthermore, platform-independent implementation in Java allows the software to run on arbitrary operating systems without the need for manual installation. The latest release of the Fit3D command line version is available at:

github.com/fkaiserbio/fit3d/releases

B.1.1. REQUIREMENTS

In order to run the command line version of Fit3D an installation of the Java Runtime Environment 1.8 or later is required. Optionally, R version 3.4.x or later is necessary to calculate *p*-values of reported matches if the FOFANOV ET AL. statistical model [161] is used. Furthermore, the user must have permissions for the installation of additional R packages or the `sfsmisc` package preinstalled. If statistical significance is estimated via the STARK ET AL. approach [162], no additional dependencies are required.

B.1.2. TEMPLATE-BASED DETECTION

In order to run a template-based structural motif detection with Fit3D the user has to specify a template motif in PDB format, the PDB identifier of a target protein, or a list of PDB identifiers for multiple targets. Alternatively, the automatic extraction of a structural motif is supported. The following simple command searches for the catalytic triad of serine proteases, extracted from the structure PDB:1gl0 provided in PDB format as file `1gl0.pdb`, in a set of PDB structures provided as list of PDB identifiers in the file `targets.txt` (all input files are located in the current path):

```
java -jar Fit3D.jar template-based -X E-H57_E-D102_E-S195 -m 1gl0.pdb -l
↪ targets.txt
```

Output When run in template-based mode, Fit3D will output matches similar to the template motif in respect to the defined RMSD upper bound (default: 2.00 Å). If no additional options (see Appendix B.1.4) are specified, the results will be written to the standard output in CSV format.

B.1.3. TEMPLATE-FREE DETECTION

To run a template-free structural motif detection with Fit3D the user needs to specify a target PDB chain which is used to get similar structures from the PDB REST API[15], a list of PDB chains, or a local directory that contains structures in PDB format. Additionally, an output directory has to be specified where results will be written. The following command detects geometrically similar structural motifs in plastocyanin structures and writes the results to the directory `results/` in the current path:

```
java -jar Fit3D.jar template-free -t 1gy2.A -o results/
```

[15] see rcsb.org/pdb/software/rest.do, available as of May 18, 2018

In this command, structure PDB:1gy2 chain A is used as a reference structure and structures according to 70% sequence similarity are automatically retrieved from the PDB REST API. Subsequently, this set of structures is used as input dataset for template-free detection.

Output When run in template-free mode, Fit3D will write the identified structural motifs clustered by structural similarity according to the consensus metric (see Section 6.2.2), in PDB format to the specified output directory. Additionally, the coverage of geometrically conserved structural motifs is written to a reference structure, encoded in B-factors. This structure can be conveniently visualized in PyMOL [147] by loading the provided `pml` script file.

B.1.4. COMMAND LINE OPTIONS

The Fit3D command line software offers a variety of advanced options to customize the structural motif detection. Table B.1 provides an overview of these options.

Table B.1.: The options for the command line version of Fit3D available in template-based (**TB**), template-free (**TF**), or both (**BO**) modes.

	short	long	description	default
BO	-a	--atoms	The identifiers of atoms that should be used to represent amino acids according to the PDB nomenclature. Identifiers must be separated by comma, e.g. 'N,CA,C,O' to use all backbone atoms. Conflicts with '-R'.	no hydrogen
	-F	--mmtf	Enables fast MMTF parsing of structures. If a local PDB installation is specified, it must contain MMTF structures. This is not compatible with the '-i' flag in template-free mode	false
	-h	--help	Displays the help dialog and terminates the application.	
	-p	--pdb	Path to a local PDB installation. Structures in this folder must be stored according to the PDB standard hierarchy: data/structures/divided/pdb/ac/pdb1acj.ent.gz for PDB and data/structures/divided/mmtf/ac/1acj.mmtf.gz for MMTF.	none
	-R	--scheme	The scheme to represent amino acids. Must be one of: 'CA' (alpha carbon), 'CB' (beta carbon), 'CO' (centroid), 'LH' (last heavy side chain atom), 'SC' (centroid of side chain atoms). Conflicts with '-a'.	none
TB	-d	--distance-tolerance	Allowed distance tolerance in Å for the extraction of local environments based on the spatial extent of the template motif.	1.00 Å
	-e	--exchange-residues	The definition of PSEs allowed for matching against the template motif. The syntax is '[number]:[type],...' For example, '12:AW,43:P' allows the template motif residue 12 to be matches against alanine and tryptophan. Residue 43 is allowed to be matched against proline.	none
	-f	--result-file	Specifies the path to the result file that will be written in CSV format.	none
	-l	--target-list	A simple text file that contains target structures separated by line break. This file may either contain entries in the format '[PDB-ID]', '[PDB-ID].[chain ID]', or paths to structures in PDB format.	none
	-m	--motif	Path to the template motif in PDB format.	none
	-n	--num-threads	Number of threads used for the calculation.	maximum
	-M	--pfam-mapping	Enables the mapping of Pfam identifiers of matches via the SIFTS [176] project. Requires Internet access.	false
	-P	--p-values	Calculation of *p*-values for matches according to [161] or [162]. Argument must be either 'F' or 'S'.	none
	-r	--rmsd	The upper bound of the RMSD up to which matches should be reported.	2.00 Å
	-t	--target	A single target structure used for detection of the template motif. Can be either [PDB-ID, [PDB-ID].[chain ID], or a path to a structure in PDB format.	none
	-U	--uniprot-mapping	Enables the mapping of UniProt identifiers of matches via the SIFTS project. Requires Internet access.	false
	-X	--extract	Extracts the motif from the input structure specified with the '-m' option ans performs a subsequent detection. Follows the syntax: '[chain]-[type][number]_...' (e.g. the three residue motif E-H57_E-D102_E-S195).	none
TF	-c	--config	Path to a user-defined configuration file in JSON format for the template-free detection algorithm. This is only recommended for expert users, please see [177] for details on the parameters.	none
	-d	--target-structures	Path to a directory that contains the target structures that should be used for detection.	none
	-i	--interactions	Enables the annotation of noncovalent inter-residue interactions with PLIP [167]. Requires Internet access and conflicts with '-F'.	false
	-l	--target-chain-list	A simple text file that contains the specification of target structures separated by line break. This file must contain entries in the format '[PDB-ID].[chain ID]'.	none
	-m	--mapping	Use a mapping scheme to group residues. Must be either 'C' (chemical groups [42]) or 'F' (functional groups).	none
	-n	--reference-chain	The reference chain that is used to visualize the coverage of geometrically conserved structural motifs. Must follow the format '[PDB-ID].[chain ID]'.	first target
	-o	--output-directory	Path to a directory where all results will be written.	none
	-r	--representative-level	Level of % sequence similarity used for the automatic retrieval of representative structures via PDB REST services. This is only used if a single target chain is specified with '-t'. Must be one of: 100, 95, 90, 70, 50, 40, or 30. Requires Internet access.	70%
	-t	--target-chain	The target chain used for template-free structural motif detection. Similar structures are automatically retrieved from via PDB REST services. Requires Internet access.	none

B.2. WEB SERVER VERSION

The web version of Fit3D supports the template-based detection of structural motifs. It was implemented using JavaServer Faces version 2.2 and PrimeFaces version 6.1. The application runs on a Tomcat 8 application server and is available at:

biosciences.hs-mittweida.de/fit3d

Input The web server version of Fit3D aims to be user-friendly and easy-to-use. It guides the user through the whole process of structural motif detection, starting with the definition of the template motif, the adjustment of parameters, and the interactive visualization of the results. Figure B.1 shows screenshots of the web server interface at different stages of template-based detection. Starting with the specification of a protein structure that contains a template motif (Figure B.1A), the user is guided through the selection of motif residues (Figure B.1B). The selected structural motif is highlighted in the structure of origin as well as displayed in its isolated form and metadata (e.g. the spatial extent) are given. Subsequently, the defined template can be directly submitted for detection (Figure B.1C). All relevant parameters, such as the RMSD upper bound or PSEs can be defined individually. The results page (Figure B.1D) shows relevant information for each individual match. The matches are sorted by ascending RMSD and annotated with additional information such as associated Pfam [132] identifiers mapped via the SIFTS project [176]. Interactive alignment visualizations of all matches (or a single match) versus the template and of global structures based on a match are found in the right panel of the application. Furthermore, the distribution of all RMSD values of the matches is given, which can be an important signature pattern for structural motifs.

Limitations Due to limited capacity of the web server, it accepts submissions of template motifs up to a size of five residues and a spatial extent up to 15.00 Å. Submissions are kept on the server for 72 hours after the calculation has finished.

B.3. API VERSION

For advanced users the API version of Fit3D is the method of choice. Template-based detection is directly integrated into the SiNGA framework [170].

B.3.1. REQUIREMENTS

The use of SiNGA requires Java Development Kit 1.8 or later to be installed. The use of the Apache Maven[16] build system is recommended. In order to use the API version of Fit3D in a custom Java project, the recommended way is to import the required Maven dependency directly from the Maven Central Repository:

```
<dependency>
        <groupId>de.bioforscher.singa</groupId>
        <artifactId>singa-structure</artifactId>
        <version>0.3.3</version>
</dependency>
```

[16] available at: maven.apache.org

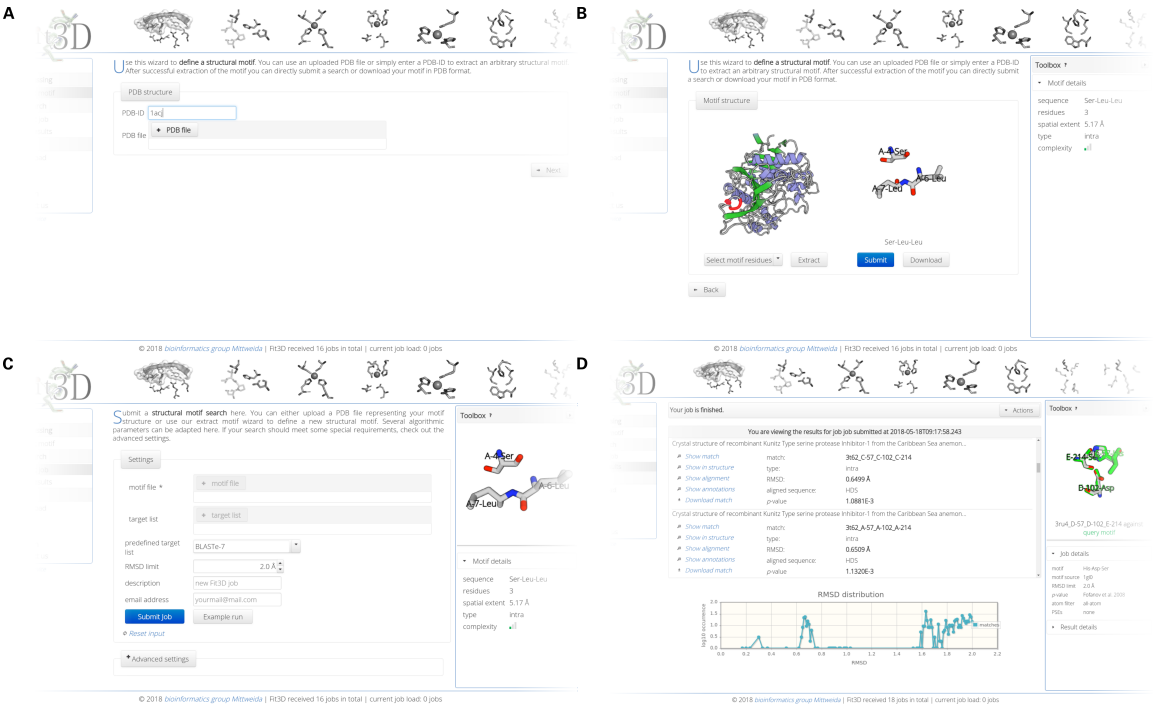


Figure B.1.: The interface of the Fit3D web sever. (**A**) The assistant for template definition allows to specify a PDB identifier or to upload a structure file in PDB format. (**B**) The structural motif is defined by selecting individual residues of the input structure. These are then interactively visualized [204] in the structure of origin and as isolated motif. Additional information about the template motif, e.g. its spatial extent, are provided in the right panel of the application. (**C**) The motif can be directly submitted for detection in the PDB or user-defined target datasets. Parameters of the Fit3D algorithm can be adjusted. The definition of PSEs is available under advanced options. (**D**) The results of a detection with Fit3D. All matches are sorted in respect to ascending RMSD values, contain metadata annotated via the SIFTS project [176], and can be displayed in the interactive structure viewer. Additionally, the distribution of all RMSD values of the current run is depicted.

B.3.2. TEMPLATE-BASED DETECTION

The following code snippet will run a template-based structural motif detection of a three-residue template motif from PDB:4cha (the catalytic triad of serine proteases) in a dataset of protein chains defined in the file `targets.txt` with PDB and chain identifiers separated by a single dot. The template consists of the residues at position 57 and 102 in chain B and residue 195 in chain C:

```
// parse the motif-containing structure
Structure structure =
 ↪  StructureParser.pdb().pdbIdentifier("4cha").everything().parse();
// define the template
StructuralMotif template = StructuralMotif.fromLeafIdentifiers(structure,
 ↪  LeafIdentifiers.of("B-57", "B-102", "C-195"));
// create a parser for multiple target structures
MultiParser multiParser =
 ↪  StructureParser.pdb().chainList(Paths.get("targets.txt"), "\\.");
// run the detection in parallel
Fit3D fit3d = Fit3DBuilder.create().query(template).targets(multiParser).
 ↪  maximalParallelism().run();
// get the matches
List<Fit3DMatch> matches = fit3d.getMatches();
```

The SiNGA documentation[17] contains further examples on how to run a template-based structural motif detection with Fit3D.

B.3.3. TEMPLATE-FREE DETECTION

Template-free detection requires algorithms that are not part of the SiNGA framework. Hence, an API for template-free structural motif detection is available as own project at:

github.com/fkaiserbio/mmm

Please note that the template-free detection API is in an early stage of development and may undergo frequent changes in due course. Figure B.2 shows the Java class diagram of evaluation metrics in the current implementation. The class model allows for an easy addition of new metrics to evaluate candidates during itemset mining. Some metrics can be evaluated in parallel (they implement the `ParallelizableMetric` interface), e.g. the extraction of candidates with cohesion is done in parallel for all structures in the dataset.

[17] available at: github.com/cleberecht/singa/wiki/Structure-Alignments-(Chemistry)

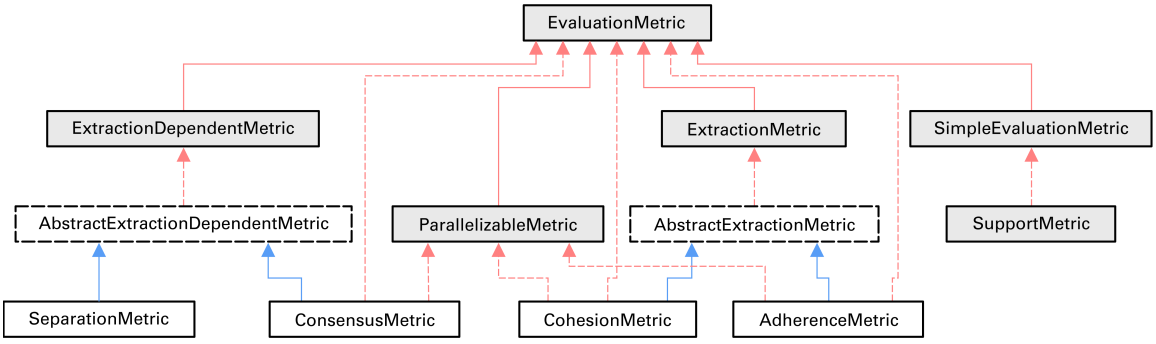


Figure B.2.: The Java class diagram for evaluation metrics in the current implementation of template-free structural motif detection. Interfaces are shaded in gray, abstract classes are depicted by dashed borders, and implementing classes are indicated by solid borders. Interface implementations are depicted by red dashed lines, interface extensions by solid red lines, and implementations or extensions by solid blue lines. The class architecture allows an easy definition of new evaluation metrics that fit into one of the metric categories simple evaluation metric, extraction metric, or extraction-dependent metric.

BIBLIOGRAPHY

[1] S. K. Burley, H. M. Berman, C. Christie, J. M. Duarte, Z. Feng, J. Westbrook, J. Young, and C. Zardecki, "RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education," *Protein Sci.*, vol. 27, pp. 316–330, Jan 2018.

[2] L. Hedstrom, "Serine protease mechanism and specificity," *Chem. Rev.*, vol. 102, pp. 4501–4524, Dec 2002.

[3] Z. Zheng and M. I. Diamond, "Huntington disease and the huntingtin protein," *Prog Mol Biol Transl Sci*, vol. 107, pp. 189–214, 2012.

[4] R. J. Najmanovich, "Evolutionary studies of ligand binding sites in proteins," *Curr. Opin. Struct. Biol.*, vol. 45, pp. 85–90, Aug 2017.

[5] I. Samish, P. E. Bourne, and R. J. Najmanovich, "Achievements and challenges in structural bioinformatics and computational biophysics," *Bioinformatics*, vol. 31, pp. 146–150, Jan 2015.

[6] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, "A minimal sequence code for switching protein structure and function," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, pp. 21149–21154, Dec 2009.

[7] T. Mukai, M. J. Lajoie, M. Englert, and D. Söll, "Rewriting the Genetic Code," *Annu. Rev. Microbiol.*, vol. 71, pp. 557–577, Sep 2017.

[8] J. H. Lee, S. K. Choi, A. Roll-Mecak, S. K. Burley, and T. E. Dever, "Universal conservation in translation initiation revealed by human and archaeal homologs of bacterial translation initiation factor IF2," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, pp. 4342–4347, Apr 1999.

[9] G. E. Fox, "Origin and evolution of the ribosome," *Cold Spring Harb Perspect Biol*, vol. 2, p. a003483, Sep 2010.

[10] M. Kafri, E. Metzl-Raz, G. Jona, and N. Barkai, "The Cost of Protein Production," *Cell Rep*, vol. 14, pp. 22–31, Jan 2016.

[11] M. Ibba and D. Söll, "Aminoacyl-tRNA synthesis," *Annu. Rev. Biochem.*, vol. 69, pp. 617–650, 2000.

[12] H. S. Bernhardt, "The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)(a)," *Biol. Direct*, vol. 7, p. 23, Jul 2012.

[13] M. Di Giulio, "The origin of the genetic code: theories and their relationships, a review," *Biosystems*, vol. 80, no. 2, pp. 175–184, 2005.

[14] W. Gilbert, "Origin of life: The RNA world," *Nature*, vol. 319, no. 6055, 1986.

[15] P. R. Wills, "The generation of meaningful information in molecular systems," *Phil. Trans. R. Soc. A*, vol. 374, no. 2063, p. 20150066, 2016.

[16] C. G. Kurland, "The RNA dreamtime: modern cells feature proteins that might have supported a prebiotic polypeptide world but nothing indicates that RNA world ever was," *Bioessays*, vol. 32, pp. 866–871, Oct 2010.

[17] R. Egel, "Peptide-dominated membranes preceding the genetic takeover by RNA: latest thinking on a classic controversy," *Bioessays*, vol. 31, pp. 1100–1109, Oct 2009.

[18] S. N. Rodin and S. Ohno, "Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid," *Orig Life Evol Biosph*, vol. 25, pp. 565–589, Dec 1995.

[19] L. Martinez-Rodriguez, O. Erdogan, M. Jimenez-Rodriguez, K. Gonzalez-Rivera, T. Williams, L. Li, V. Weinreb, M. Collier, S. N. Chandrasekaran, X. Ambroggio, B. Kuhlman, and C. W. Carter, "Functional Class I and II Amino Acid-activating Enzymes Can Be Coded by Opposite Strands of the Same Gene," *J. Biol. Chem.*, vol. 290, pp. 19710–19725, Aug 2015.

[20] P. R. Wills, "Spontaneous mutual ordering of nucleic acids and proteins," *Orig Life Evol Biosph*, vol. 44, pp. 293–298, Dec 2014.

[21] A. Harish and G. Caetano-Anolles, "Ribosomal history reveals origins of modern protein synthesis," *PLoS ONE*, vol. 7, no. 3, p. e32776, 2012.

[22] C. W. Carter, "Coding of Class I and II Aminoacyl-tRNA Synthetases," *Adv. Exp. Med. Biol.*, vol. 966, pp. 103–148, 2017.

[23] P. R. Wills and C. W. Carter, "Insuperable problems of the genetic code initially emerging in an RNA world," *BioSystems*, vol. 164, pp. 155–166, Feb 2018.

[24] C. W. Carter and P. R. Wills, "Interdependence, Reflexivity, Fidelity, Impedance Matching, and the Evolution of Genetic Coding," *Mol. Biol. Evol.*, vol. 35, pp. 269–286, Feb 2018.

[25] C. W. Carter and R. Wolfenden, "tRNA acceptor stem and anticodon bases form independent codes related to protein folding," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, pp. 7489–7494, Jun 2015.

[26] A. Dock-Bregeon, R. Sankaranarayanan, P. Romby, J. Caillet, M. Springer, B. Rees, C. S. Francklyn, C. Ehresmann, and D. Moras, "Transfer RNA-mediated editing in threonyl-tRNA synthetase. The class II solution to the double discrimination problem," *Cell*, vol. 103, pp. 877–884, Dec 2000.

[27] A. Hadd and J. J. Perona, "Coevolution of specificity determinants in eukaryotic glutamyl- and glutaminyl-tRNA synthetases," *J. Mol. Biol.*, vol. 426, pp. 3619–3633, Oct 2014.

[28] N. Nair, H. Raff, M. T. Islam, M. Feen, D. M. Garofalo, and K. Sheppard, "The Bacillus subtilis and Bacillus halodurans Aspartyl-tRNA Synthetases Retain Recognition of tRNA(Asn)," *J. Mol. Biol.*, vol. 428, pp. 618–630, Feb 2016.

[29] Y. Song, H. Zhou, M. N. Vo, Y. Shi, M. H. Nawaz, O. Vargas-Rodriguez, J. K. Diedrich, J. R. Yates, S. Kishi, K. Musier-Forsyth, and P. Schimmel, "Double mimicry evades tRNA synthetase editing by toxic vegetable-sourced non-proteinogenic amino acid," *Nat Commun*, vol. 8, p. 2281, Dec 2017.

[30] A. Koutsotoli and A. G. Tzakos, "Host-pathogen crosstalking: the mastery of taking the helm of the host," *Structure*, vol. 20, pp. 1613–1615, Oct 2012.

[31] J. P. Nilmeier, D. A. Kirshner, S. E. Wong, and F. C. Lightstone, "Rapid catalytic template searching as an enzyme function prediction procedure," *PLoS ONE*, vol. 8, no. 5, p. e62535, 2013.

[32] Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang, and L. Chen, "Bridging protein local structures and protein functions," *Amino Acids*, vol. 35, pp. 627–650, Oct 2008.

[33] E. C. Meng, B. J. Polacco, and P. C. Babbitt, "Superfamily active site templates," *Proteins*, vol. 55, pp. 962–976, Jun 2004.

[34] B. Schmidt and A. Hildebrandt, "Next-generation sequencing: big data meets high performance computing," *Drug Discov. Today*, vol. 22, pp. 712–717, Apr 2017.

[35] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model," *PLoS Comput. Biol.*, vol. 13, p. e1005324, Jan 2017.

[36] K. Uziela, D. Menendez Hurtado, N. Shu, B. Wallner, and A. Elofsson, "ProQ3D: improved model quality assessments using deep learning," *Bioinformatics*, vol. 33, pp. 1578–1580, May 2017.

[37] I. Wallach, M. Dzamba, and A. Heifets, "Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery," *CoRR*, vol. abs/1510.02855, 2015.

[38] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235–242, Jan 2000.

[39] P. W. Rose, A. Prlic, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y. P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. Yang, J. Y. Young, C. Zardecki, H. M. Berman, and S. K. Burley, "The RCSB protein data bank: integrative view of protein, gene and 3D structural information," *Nucleic Acids Res.*, vol. 45, pp. D271–D281, Jan 2017.

[40] S. Ovchinnikov, H. Park, N. Varghese, P. S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker, "Protein structure determination using metagenome sequence data," *Science*, vol. 355, pp. 294–298, Jan 2017.

[41] E. Webb, *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Published for the International Union of Biochemistry and Molecular Biology by Academic Press, 1992.

[42] A. Gutteridge and J. M. Thornton, "Understanding nature's catalytic toolkit," *Trends in biochemical sciences*, vol. 30, no. 11, pp. 622–629, 2005.

[43] S. C. Izidoro, R. C. de Melo-Minardi, and G. L. Pappa, "GASS: identifying enzyme active sites with genetic algorithms," *Bioinformatics*, vol. 31, pp. 864–870, Mar 2015.

[44] M. Brylinski, "eMatchSite: sequence order-independent structure alignments of ligand binding pockets in protein models," *PLoS Comput. Biol.*, vol. 10, p. e1003829, Sep 2014.

[45] Z. Zhao, L. Xie, and P. E. Bourne, "Insights into the binding mode of MEK type-III inhibitors. A step towards discovering and designing allosteric kinase inhibitors across the human kinome," *PLoS ONE*, vol. 12, no. 6, p. e0179936, 2017.

[46] V. Bonnici, R. Giugno, A. Pulvirenti, D. Shasha, and A. Ferro, "A subgraph isomorphism algorithm and its application to biochemical data," *BMC Bioinformatics*, vol. 14 Suppl 7, p. S13, 2013.

[47] G. Gonzalez, B. Hannigan, and W. F. DeGrado, "A real-time all-atom structural search engine for proteins," *PLoS Comput. Biol.*, vol. 10, p. e1003750, Jul 2014.

[48] J. T. Wong, "A co-evolution theory of the genetic code," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 72, pp. 1909–1912, May 1975.

[49] T. Sonneborn, "Degeneracy of the genetic code: extent, nature, and genetic implications," *Evolving genes and proteins*, pp. 377–397, 1965.

[50] C. R. Woese, "Order in the genetic code," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 54, pp. 71–75, Jul 1965.

[51] R. C. Guimaraes, C. H. Moreira, and S. T. de Farias, "A self-referential model for the formation of the genetic code," *Theory Biosci.*, vol. 127, pp. 249–270, Aug 2008.

[52] J. T. Wong, "Coevolution theory of the genetic code at age thirty," *Bioessays*, vol. 27, pp. 416–425, Apr 2005.

[53] J. G. Arnez and D. Moras, "Structural and functional considerations of the aminoacylation reaction," *Trends Biochem. Sci.*, vol. 22, pp. 211–216, Jun 1997.

[54] M. Praetorius-Ibba, N. Stange-Thomann, M. Kitabatake, K. Ali, I. Söll, C. W. Carter, M. Ibba, and D. Söll, "Ancient adaptation of the active site of tryptophanyl-tRNA synthetase for tryptophan binding," *Biochemistry*, vol. 39, pp. 13136–13143, Oct 2000.

[55] J. J. Burbaum and P. Schimmel, "Structural relationships and the classification of aminoacyl-tRNA synthetases," *J. Biol. Chem.*, vol. 266, pp. 16965–16968, Sep 1991.

[56] L. Ribas de Pouplana and P. Schimmel, "Aminoacyl-tRNA synthetases: potential markers of genetic code development," *Trends Biochem. Sci.*, vol. 26, pp. 591–596, Oct 2001.

[57] P. Schimmel and T. Ripmaster, "Modular design of components of the operational RNA code for alanine in evolution," *Trends Biochem. Sci.*, vol. 20, pp. 333–334, Sep 1995.

[58] Y. I. Wolf, L. Aravind, N. V. Grishin, and E. V. Koonin, "Evolution of aminoacyl-tRNA synthetases – analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events," *Genome Res.*, vol. 9, pp. 689–710, Aug 1999.

[59] A. Chaliotis, P. Vlastaridis, D. Mossialos, M. Ibba, H. D. Becker, C. Stathopoulos, and G. D. Amoutzias, "The complex evolutionary history of aminoacyl-tRNA synthetases," *Nucleic Acids Res.*, vol. 45, pp. 1059–1068, Feb 2017.

[60] M. A. Rould, J. J. Perona, and T. A. Steitz, "Structural basis of anticodon loop recognition by glutaminyl-tRNA synthetase," *Nature*, vol. 352, pp. 213–218, Jul 1991.

[61] J. Normanly and J. Abelson, "tRNA identity," *Annu. Rev. Biochem.*, vol. 58, pp. 1029–1049, 1989.

[62] S. A. Martinis and M. T. Boniecki, "The balance between pre- and post-transfer editing in tRNA synthetases," *FEBS Lett.*, vol. 584, pp. 455–459, Jan 2010.

[63] K. E. Splan, M. E. Ignatov, and K. Musier-Forsyth, "Transfer RNA modulates the editing mechanism used by class II prolyl-tRNA synthetase," *J. Biol. Chem.*, vol. 283, pp. 7128–7134, Mar 2008.

[64] G. Eriani, M. Delarue, O. Poch, J. Gangloff, and D. Moras, "Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs," *Nature*, vol. 347, pp. 203–206, Sep 1990.

[65] D. Moras, "Structural and functional relationships between aminoacyl-tRNA synthetases," *Trends Biochem. Sci.*, vol. 17, pp. 159–164, Apr 1992.

[66] C. D. Livingstone and G. J. Barton, "Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation," *Comput. Appl. Biosci.*, vol. 9, pp. 745–756, Dec 1993.

[67] H. Belrhali, A. Yaremchuk, M. Tukalo, C. Berthet-Colominas, B. Rasmussen, P. Bosecke, O. Diat, and S. Cusack, "The structural basis for seryl-adenylate and Ap4A synthesis by seryl-tRNA synthetase," *Structure*, vol. 3, pp. 341–352, Apr 1995.

[68] M. Fujinaga, C. Berthet-Colominas, A. D. Yaremchuk, M. A. Tukalo, and S. Cusack, "Refined crystal structure of the seryl-tRNA synthetase from Thermus thermophilus at 2.5 A resolution," *J. Mol. Biol.*, vol. 234, pp. 222–233, Nov 1993.

[69] A. Ambrogelly, D. Söll, O. Nureki, S. Yokoyama, and M. Ibba, *Class I Lysyl-tRNA Synthetases*. Landes Bioscience, 2013.

[70] Y. Diaz-Lazcoz, J. C. Aude, P. Nitschke, H. Chiapello, C. Landes-Devauchelle, and J. L. Risler, "Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases," *Mol. Biol. Evol.*, vol. 15, pp. 1548–1561, Nov 1998.

[71] C. R. Woese, G. J. Olsen, M. Ibba, and D. Söll, "Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process," *Microbiol. Mol. Biol. Rev.*, vol. 64, pp. 202–236, Mar 2000.

[72] E. Schmitt, M. Panvert, S. Blanquet, and Y. Mechulam, "Transition state stabilization by the 'HIGH' motif of class I aminoacyl-tRNA synthetases: the case of Escherichia coli methionyl-tRNA synthetase," *Nucleic Acids Res.*, vol. 23, pp. 4793–4798, Dec 1995.

[73] E. A. First and A. R. Fersht, "Involvement of threonine 234 in catalysis of tyrosyl adenylate formation by tyrosyl-tRNA synthetase," *Biochemistry*, vol. 32, pp. 13644–13650, Dec 1993.

[74] E. A. First and A. R. Fersht, "Mutation of lysine 233 to alanine introduces positive cooperativity into tyrosyl-tRNA synthetase," *Biochemistry*, vol. 32, pp. 13651–13657, Dec 1993.

[75] E. A. First and A. R. Fersht, "Mutational and kinetic analysis of a mobile loop in tyrosyl-tRNA synthetase," *Biochemistry*, vol. 32, pp. 13658–13663, Dec 1993.

[76] E. A. First and A. R. Fersht, "Analysis of the role of the KMSKS loop in the catalytic mechanism of the tyrosyl-tRNA synthetase using multimutant cycles," *Biochemistry*, vol. 34, pp. 5030–5043, Apr 1995.

[77] S. N. Chandrasekaran, J. Das, N. V. Dokholyan, and C. W. Carter, "A modified PATH algorithm rapidly generates transition states comparable to those found by other well established algorithms," *Struct Dyn*, vol. 3, p. 012101, Jan 2016.

[78] S. N. Chandrasekaran and C. W. Carter, "Augmenting the anisotropic network model with torsional potentials improves PATH performance, enabling detailed comparison with experimental rate data," *Struct Dyn*, vol. 4, p. 032103, May 2017.

[79] C. W. Carter, S. N. Chandrasekaran, V. Weinreb, L. Li, and T. Williams, "Combining multi-mutant and modular thermodynamic cycles to measure energetic coupling networks in enzyme catalysis," *Struct Dyn*, vol. 4, p. 032101, May 2017.

[80] V. Weinreb, L. Li, and C. W. Carter, "A master switch couples Mg2+-assisted catalysis to domain motion in B. stearothermophilus tryptophanyl-tRNA Synthetase," *Structure*, vol. 20, pp. 128–138, Jan 2012.

[81] V. Weinreb, L. Li, S. N. Chandrasekaran, P. Koehl, M. Delarue, and C. W. Carter, "Enhanced amino acid selection in fully evolved tryptophanyl-tRNA synthetase, relative to its Urzyme, requires domain motion sensed by the D1 switch, a remote dynamic packing motif," *J. Biol. Chem.*, vol. 289, pp. 4367–4376, Feb 2014.

[82] G. Eriani, J. Cavarelli, F. Martin, G. Dirheimer, D. Moras, and J. Gangloff, "Role of dimerization in yeast aspartyl-tRNA synthetase and importance of the class II invariant proline," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 90, pp. 10816–10820, Nov 1993.

[83] A. Aberg, A. Yaremchuk, M. Tukalo, B. Rasmussen, and S. Cusack, "Crystal structure analysis of the activation of histidine by Thermus thermophilus histidyl-tRNA synthetase," *Biochemistry*, vol. 36, pp. 3084–3094, Mar 1997.

[84] S. Cusack, "Aminoacyl-tRNA synthetases," *Curr. Opin. Struct. Biol.*, vol. 7, pp. 881–889, Dec 1997.

[85] S. Cusack, C. Berthet-Colominas, M. Hartlein, N. Nassar, and R. Leberman, "A second class of synthetase structure revealed by X-ray analysis of Escherichia coli seryl-tRNA synthetase at 2.5 Å," *Nature*, vol. 347, pp. 249–255, Sep 1990.

[86] S. Cusack, M. Hartlein, and R. Leberman, "Sequence, structural and evolutionary relationships between class 2 aminoacyl-tRNA synthetases," *Nucleic Acids Res.*, vol. 19, pp. 3489–3498, Jul 1991.

[87] P. O'Donoghue and Z. Luthey-Schulten, "On the evolution of structure in aminoacyl-tRNA synthetases," *Microbiol. Mol. Biol. Rev.*, vol. 67, pp. 550–573, Dec 2003.

[88] S. D. Banik and N. Nandi, "Mechanism of the activation step of the aminoacylation reaction: a significant difference between class I and class II synthetases," *J. Biomol. Struct. Dyn.*, vol. 30, no. 6, pp. 701–715, 2012.

[89] J. R. Brown and W. F. Doolittle, "Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 92, pp. 2441–2445, Mar 1995.

[90] P. Schimmel, R. Giege, D. Moras, and S. Yokoyama, "An operational RNA code for amino acids and possible relationship to genetic code," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 90, pp. 8763–8768, Oct 1993.

[91] S. N. Chandrasekaran, G. G. Yardimci, O. Erdogan, J. Roach, and C. W. Carter, "Statistical evaluation of the Rodin-Ohno hypothesis: sense/antisense coding of ancestral class I and II aminoacyl-tRNA synthetases," *Mol. Biol. Evol.*, vol. 30, pp. 1588–1604, Jul 2013.

[92] H. B. LeJohn, L. E. Cameron, B. Yang, and S. L. Rennie, "Molecular characterization of an NAD-specific glutamate dehydrogenase gene inducible by L-glutamine. Antisense gene pair arrangement with L-glutamine-inducible heat shock 70-like protein gene," *J. Biol. Chem.*, vol. 269, pp. 4523–4531, Feb 1994.

[93] C. W. Carter and W. L. Duax, "Did tRNA synthetase classes arise on opposite strands of the same gene?," *Mol. Cell*, vol. 10, pp. 705–708, Oct 2002.

[94] J. Chen, M. Sun, W. J. Kent, X. Huang, H. Xie, W. Wang, G. Zhou, R. Z. Shi, and J. D. Rowley, "Over 20% of human transcripts might form sense-antisense pairs," *Nucleic Acids Res.*, vol. 32, no. 16, pp. 4812–4820, 2004.

[95] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C. C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K. C. Pang, J. Hallinan, J. Mattick, D. A. Hume, L. Lipovich, S. Batalov, P. G. Engstrom, Y. Mizuno, M. A. Faghihi, A. Sandelin, A. M. Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard, and C. Wahlestedt, "Antisense transcription in the mammalian transcriptome," *Science*, vol. 309, pp. 1564–1566, Sep 2005.

[96] M. Ruff, S. Krishnaswamy, M. Boeglin, A. Poterszman, A. Mitschler, A. Podjarny, B. Rees, J. C. Thierry, and D. Moras, "Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp)," *Science*, vol. 252, pp. 1682–1689, Jun 1991.

[97] C. W. Carter, L. Li, V. Weinreb, M. Collier, K. Gonzalez-Rivera, M. Jimenez-Rodriguez, O. Erdogan, B. Kuhlman, X. Ambroggio, T. Williams, and S. N. Chandrasekharan, "The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed," *Biol. Direct*, vol. 9, p. 11, Jun 2014.

[98] J. O. Schulze, A. Masoumi, D. Nickel, M. Jahn, D. Jahn, W. D. Schubert, and D. W. Heinz, "Crystal structure of a non-discriminating glutamyl-tRNA synthetase," *J. Mol. Biol.*, vol. 361, pp. 888–897, Sep 2006.

[99] B. M. Mailu, G. Ramasamay, D. G. Mudeppa, L. Li, S. E. Lindner, M. J. Peterson, A. E. DeRocher, S. H. Kappe, P. K. Rathod, and M. J. Gardner, "A nondiscriminating glutamyl-tRNA synthetase in the plasmodium apicoplast: the first enzyme in an indirect aminoacylation pathway," *J. Biol. Chem.*, vol. 288, pp. 32539–32552, Nov 2013.

[100] S. Dutta, K. Choudhury, S. D. Banik, and N. Nandi, "Active site nanospace of aminoacyl tRNA synthetase: difference between the class I and class II synthetases," *J Nanosci Nanotechnol*, vol. 14, pp. 2280–2298, Mar 2014.

[101] Y. Pham, L. Li, A. Kim, O. Erdogan, V. Weinreb, G. L. Butterfoss, B. Kuhlman, and C. W. Carter, "A minimal TrpRS catalytic domain supports sense/antisense ancestry of class I and II aminoacyl-tRNA synthetases," *Mol. Cell*, vol. 25, pp. 851–862, Mar 2007.

[102] J. Hou, S. R. Jun, C. Zhang, and S. H. Kim, "Global mapping of the protein structure space and application in structure-based inference of protein function," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 3651–3656, Mar 2005.

[103] I. Jonassen, I. Eidhammer, and W. R. Taylor, "Discovery of local packing motifs in protein structures," *Proteins*, vol. 34, pp. 206–219, Feb 1999.

[104] M. Dudev and C. Lim, "Discovering structural motifs using a structural alphabet: application to magnesium-binding sites," *BMC bioinformatics*, vol. 8, no. 1, p. 1, 2007.

[105] I. Jonassen, I. Eidhammer, D. Conklin, and W. R. Taylor, "Structure motif discovery and mining the PDB," *Bioinformatics*, vol. 18, no. 2, pp. 362–367, 2002.

[106] H. R. Brodkin, N. A. DeLateur, S. Somarowthu, C. L. Mills, W. R. Novak, P. J. Beuning, D. Ringe, and M. J. Ondrechen, "Prediction of distal residue participation in enzyme catalysis," *Protein Sci.*, vol. 24, pp. 762–778, May 2015.

[107] P. Colman, H. Freeman, J. Guss, M. Murata, V. Norris, J. Ramshaw, and M. Venkatappa, "X-ray crystal structure analysis of plastocyanin at 2.7 Å resolution," *Nature*, vol. 272, pp. 319–324, 1978.

[108] J. Coloma, R. Jain, K. R. Rajashankar, A. Garcia-Sastre, and A. K. Aggarwal, "Structures of NS5 Methyltransferase from Zika Virus," *Cell Rep*, vol. 16, pp. 3097–3102, Sep 2016.

[109] J. Miller, A. D. McLachlan, and A. Klug, "Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes," *EMBO J.*, vol. 4, pp. 1609–1614, Jun 1985.

[110] R. B. Darnell, "Developing global insight into RNA regulation," *Cold Spring Harb. Symp. Quant. Biol.*, vol. 71, pp. 321–327, 2006.

[111] G. K. Kong, J. J. Adams, H. H. Harris, J. F. Boas, C. C. Curtain, D. Galatis, C. L. Masters, K. J. Barnham, W. J. McKinstry, R. Cappai, and M. W. Parker, "Structural studies of the Alzheimer's amyloid precursor protein copper-binding domain reveal how it binds copper ions," *J. Mol. Biol.*, vol. 367, pp. 148–161, Mar 2007.

[112] K. H. Ebrahimi, P. L. Hagedoorn, and W. R. Hagen, "A synthetic peptide with the putative iron binding motif of amyloid precursor protein (APP) does not catalytically oxidize iron," *PLoS ONE*, vol. 7, no. 8, p. e40287, 2012.

[113] Y. Xue, A. V. Davis, G. Balakrishnan, J. P. Stasser, B. M. Staehlin, P. Focia, T. G. Spiro, J. E. Penner-Hahn, and T. V. O'Halloran, "Cu(I) recognition via cation-pi and methionine interactions in CusF," *Nat. Chem. Biol.*, vol. 4, pp. 107–109, Feb 2008.

[114] B. W. Matthews, P. B. Sigler, R. Henderson, and D. M. Blow, "Three-dimensional structure of tosyl-alpha-chymotrypsin," *Nature*, vol. 214, pp. 652–656, May 1967.

[115] D. Fischer, H. Wolfson, S. L. Lin, and R. Nussinov, "Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding," *Protein Sci.*, vol. 3, pp. 769–778, May 1994.

[116] M. J. Plevin, D. L. Bryce, and J. Boisbouvier, "Direct detection of CH/pi interactions in proteins," *Nat Chem*, vol. 2, pp. 466–471, Jun 2010.

[117] T. W. Craven, M. K. Cho, N. J. Traaseth, R. Bonneau, and K. Kirshenbaum, "A Miniature Protein Stabilized by a Cation-π Interaction Network," *J. Am. Chem. Soc.*, vol. 138, pp. 1543–1550, Feb 2016.

[118] Y. Valasatava, A. Rosato, N. Furnham, J. M. Thornton, and C. Andreini, "To what extent do structural changes in catalytic metal sites affect enzyme function?," *Journal of Inorganic Biochemistry*, 2017.

[119] S. S. Krishna, I. Majumdar, and N. V. Grishin, "Structural classification of zinc fingers: survey and summary," *Nucleic Acids Res.*, vol. 31, pp. 532–550, Jan 2003.

[120] P. C. Babbitt, M. S. Hasson, J. E. Wedekind, D. R. Palmer, W. C. Barrett, G. H. Reed, I. Rayment, D. Ringe, G. L. Kenyon, and J. A. Gerlt, "The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids," *Biochemistry*, vol. 35, pp. 16489–16501, Dec 1996.

[121] J. S. Fetrow and P. C. Babbitt, "New computational approaches to understanding molecular protein function," *PLoS Comput. Biol.*, vol. 14, p. e1005756, Apr 2018.

[122] G. Nunez-Vivanco, A. Valdes-Jimenez, F. Besoain, and M. Reyes-Parada, "Geomfinder: a multi-feature identifier of similar three-dimensional protein patterns: a ligand-independent approach," *J Cheminform*, vol. 8, p. 19, 2016.

[123] J. S. Fetrow and J. Skolnick, "Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases," *J. Mol. Biol.*, vol. 281, pp. 949–968, Sep 1998.

[124] N. Furnham, G. L. Holliday, T. A. de Beer, J. O. Jacobsen, W. R. Pearson, and J. M. Thornton, "The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes," *Nucleic Acids Res.*, vol. 42, pp. D485–489, Jan 2014.

[125] A. J. M. Ribeiro, G. L. Holliday, N. Furnham, J. D. Tyzack, K. Ferris, and J. M. Thornton, "Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites," *Nucleic Acids Res.*, Nov 2017.

[126] G. L. Holliday, S. D. Brown, E. Akiva, D. Mischel, M. A. Hicks, J. H. Morris, C. C. Huang, E. C. Meng, S. C. Pegg, T. E. Ferrin, and P. C. Babbitt, "Biocuration in the structure-function linkage database: the anatomy of a superfamily," *Database (Oxford)*, vol. 2017, Jan 2017.

[127] J. Konc and D. Janezic, "ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment," *Bioinformatics*, vol. 26, pp. 1160–1168, May 2010.

[128] N. Nadzirin, E. J. Gardiner, P. Willett, P. J. Artymiuk, and M. Firdaus-Raih, "SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures," *Nucleic Acids Res.*, vol. 40, pp. W380–386, Jul 2012.

[129] G. J. Kleywegt, "Recognition of spatial motifs in protein structures," *J. Mol. Biol.*, vol. 285, pp. 1887–1897, Jan 1999.

[130] G. Debret, A. Martel, and P. Cuniasse, "RASMOT-3D PRO: a 3D motif search webserver," *Nucleic Acids Res.*, vol. 37, pp. W459–464, Jul 2009.

[131] J. Huan, D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, and A. Tropsha, "Comparing graph representations of protein structure for mining family-specific residue-based packing motifs," *J. Comput. Biol.*, vol. 12, no. 6, pp. 657–671, 2005.

[132] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, "The Pfam protein families database: towards a more sustainable future," *Nucleic Acids Res.*, vol. 44, pp. D279–285, Jan 2016.

[133] F. Kaiser, A. Eisold, S. Bittrich, and D. Labudde, "Fit3D: a web application for highly accurate screening of spatial residue patterns in protein structure data," *Bioinformatics*, vol. 32, pp. 792–794, Mar 2016.

[134] V. J. Haupt and M. Schroeder, "Old friends in new guise: repositioning of known drugs with structural bioinformatics," *Brief. Bioinformatics*, vol. 12, pp. 312–326, Jul 2011.

[135] A. S. Mian, M. Bennamoun, and R. Owens, "Three-dimensional model-based object recognition and segmentation in cluttered scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1584–1601, Oct 2006.

[136] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 34, no. 5, pp. 827–828, 1978.

[137] L. He, F. Vandin, G. Pandurangan, and C. Bailey-Kellogg, "Ballast: a ball-based algorithm for structural motifs," *J. Comput. Biol.*, vol. 20, pp. 137–151, Feb 2013.

[138] G. Ausiello, A. Via, and M. Helmer-Citterich, "Query3d: a new method for high-throughput analysis of functional residues in protein structures," *BMC Bioinformatics*, vol. 6 Suppl 4, p. S5, Dec 2005.

[139] M. Gao and J. Skolnick, "APoc: large-scale identification of similar protein pockets," *Bioinformatics*, vol. 29, pp. 597–604, Mar 2013.

[140] T. A. Binkowski, P. Freeman, and J. Liang, "pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins," *Nucleic Acids Res.*, vol. 32, pp. W555–558, Jul 2004.

[141] T. A. Binkowski, L. Adamian, and J. Liang, "Inferring functional relationships of proteins from local sequence and spatial surface patterns," *J. Mol. Biol.*, vol. 332, pp. 505–526, Sep 2003.

[142] M. Moll, D. H. Bryant, and L. E. Kavraki, "The LabelHash algorithm for substructure matching," *BMC Bioinformatics*, vol. 11, p. 555, Nov 2010.

[143] M. Moll, D. H. Bryant, and L. E. Kavraki, "The LabelHash server and tools for substructure-based functional annotation," *Bioinformatics*, vol. 27, pp. 2161–2162, Aug 2011.

[144] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson, "Recognition of functional sites in protein structures," *J. Mol. Biol.*, vol. 339, pp. 607–633, Jun 2004.

[145] J. Konc and D. Janezic, "ProBiS tools (algorithm, database, and web servers) for predicting and modeling of biologically interesting proteins," *Prog. Biophys. Mol. Biol.*, vol. 128, pp. 24–32, Sep 2017.

[146] G. H. Li and J. F. Huang, "CMASA: an accurate algorithm for detecting local protein structural similarity and its application to enzyme catalytic site annotation," *BMC Bioinformatics*, vol. 11, p. 439, 2010.

[147] Schrödinger, LLC, "The PyMOL molecular graphics system, version 1.8." Nov 2015.

[148] C. C. Aggarwal, *Data mining: the textbook*. Springer, 2015.

[149] S. Naulaerts, P. Meysman, W. Bittremieux, T. N. Vu, W. Vanden Berghe, B. Goethals, and K. Laukens, "A primer to frequent itemset mining for bioinformatics," *Brief. Bioinformatics*, vol. 16, pp. 216–231, Mar 2015.

[150] C. Zhou, P. Meysman, B. Cule, K. Laukens, and B. Goethals, "Discovery of Spatially Cohesive Itemsets in Three-Dimensional Protein Structures," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 11, no. 5, pp. 814–825, 2014.

[151] P. Meysman, C. Zhou, B. Cule, B. Goethals, and K. Laukens, "Mining the entire Protein DataBank for frequent spatially cohesive amino acid patterns," *BioData Min*, vol. 8, p. 4, 2015.

[152] W. Dhifli, R. Saidi, and E. M. Nguifo, "Smoothing 3D protein structure motifs through graph mining and amino acid similarities," *J. Comput. Biol.*, vol. 21, pp. 162–172, Feb 2014.

[153] J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha, "Mining protein family specific residue packing patterns from protein structure graphs," in *Proceedings of the eighth annual international conference on Research in computational molecular biology*, pp. 308–315, ACM, 2004.

[154] L. Xie and P. E. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, pp. 5441–5446, Apr 2008.

[155] F. Guyon and P. Tuffery, "Fast protein fragment similarity scoring using a Binet-Cauchy kernel," *Bioinformatics*, vol. 30, pp. 784–791, Mar 2014.

[156] F. Guyon, F. Martz, M. Vavrusa, J. Becot, J. Rey, and P. Tuffery, "BCSearch: fast structural fragment mining over large collections of protein structures," *Nucleic Acids Res.*, vol. 43, pp. W378–382, Jul 2015.

[157] G. Ausiello, P. F. Gherardini, P. Marcatili, A. Tramontano, A. Via, and M. Helmer-Citterich, "FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures," *BMC Bioinformatics*, vol. 9 Suppl 2, p. S2, Mar 2008.

[158] A. Zemla, "LGA: A method for finding 3D similarities in protein structures," *Nucleic Acids Res.*, vol. 31, pp. 3370–3374, Jul 2003.

[159] H. Tsukada and D. M. Blow, "Structure of alpha-chymotrypsin refined at 1.68 Å resolution," *J. Mol. Biol.*, vol. 184, pp. 703–711, Aug 1985.

[160] A. Wacker, J. Buck, D. Mathieu, C. Richter, J. Wohnert, and H. Schwalbe, "Structure and dynamics of the deoxyguanosine-sensing riboswitch studied by NMR-spectroscopy," *Nucleic Acids Res.*, vol. 39, pp. 6802–6812, Aug 2011.

[161] V. Fofanov, B. Chen, D. Bryant, M. Moll, O. Lichtarge, L. Kavraki, and M. Kimmel, "A statistical model to correct systematic bias introduced by algorithmic thresholds in protein structural comparison algorithms," in *Bioinformatics and Biomedicine Workshops, 2008. BIBMW 2008. IEEE International Conference on*, pp. 1–8, Nov 2008.

[162] A. Stark, S. Sunyaev, and R. B. Russell, "A model for statistical significance of local similarities in structure," *J. Mol. Biol.*, vol. 326, pp. 1307–1316, Mar 2003.

[163] F. Kaiser, S. Bittrich, S. Salentin, C. Leberecht, V. J. Haupt, S. Krautwurst, M. Schroeder, and D. Labudde, "Backbone Brackets and Arginine Tweezers delineate Class I and Class II aminoacyl tRNA synthetases," *PLoS Comput. Biol.*, vol. 14, p. e1006101, Apr 2018.

[164] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Res.*, vol. 33, no. 7, pp. 2302–2309, 2005.

[165] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, pp. 443–453, Mar 1970.

[166] F. Armougom, S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, and C. Notredame, "Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee," *Nucleic Acids Res.*, vol. 34, pp. W604–608, Jul 2006.

[167] S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, and M. Schroeder, "PLIP: fully automated protein-ligand interaction profiler," *Nucleic Acids Res.*, vol. 43, pp. W443–447, Jul 2015.

[168] S. Salentin, V. J. Haupt, S. Daminelli, and M. Schroeder, "Polypharmacology rescored: protein-ligand interaction profiles for remote binding site similarity assessment," *Prog. Biophys. Mol. Biol.*, vol. 116, no. 2-3, pp. 174–186, 2014.

[169] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Res.*, vol. 14, pp. 1188–1190, Jun 2004.

[170] C. Leberecht, F. Kaiser, and S. Bittrich, "cleberecht/singa: singa-all release 0.3.3," May 2018.

[171] G. R. Stockwell and J. M. Thornton, "Conformational diversity of ligands bound to proteins," *J. Mol. Biol.*, vol. 356, pp. 928–944, Mar 2006.

[172] C. Ceccarelli, N. B. Grodsky, N. Ariyaratne, R. F. Colman, and B. J. Bahnson, "Crystal structure of porcine mitochondrial NADP+-dependent isocitrate dehydrogenase complexed with Mn2+ and isocitrate. Insights into the enzyme mechanism," *J. Biol. Chem.*, vol. 277, pp. 43454–43462, Nov 2002.

[173] T. Yanagisawa, T. Sumida, R. Ishii, C. Takemoto, and S. Yokoyama, "A paralog of lysyl-tRNA synthetase aminoacylates a conserved lysine residue in translation elongation factor P," *Nat. Struct. Mol. Biol.*, vol. 17, pp. 1136–1143, Sep 2010.

[174] M. Mocibob, N. Ivic, M. Luic, and I. Weygand-Durasevic, "Adaptation of aminoacyl-tRNA synthetase catalytic core to carrier protein aminoacylation," *Structure*, vol. 21, pp. 614–626, Apr 2013.

[175] T. UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res.*, vol. 46, p. 2699, Mar 2018.

[176] S. Velankar, J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M. J. Martin, and G. J. Kleywegt, "SIFTS: Structure Integration with Function, Taxonomy and Sequences resource," *Nucleic Acids Res.*, vol. 41, pp. D483–489, Jan 2013.

[177] F. Kaiser and D. Labudde, "Unsupervised Discovery of Geometrically Common Structural Motifs and Long-Range Contacts in Protein 3D Structures," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. PP, pp. 1–1, Dec 2017.

[178] V. Alva, J. Soding, and A. N. Lupas, "A vocabulary of ancient peptides at the origin of folded proteins," *Elife*, vol. 4, p. e09410, Dec 2015.

[179] Y. Zhang, O. Zagnitko, I. Rodionova, A. Osterman, and A. Godzik, "The FGGY carbohydrate kinase family: insights into the evolution of functional specificities," *PLoS Comput. Biol.*, vol. 7, p. e1002318, Dec 2011.

[180] G. L. Challis, J. Ravel, and C. A. Townsend, "Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains," *Chem. Biol.*, vol. 7, pp. 211–224, Mar 2000.

[181] G. Caetano-Anolles, M. Wang, D. Caetano-Anolles, and J. E. Mittenthal, "The origin, evolution and structure of the protein world," *Biochem. J.*, vol. 417, pp. 621–637, Feb 2009.

[182] B. Delagoutte, D. Moras, and J. Cavarelli, "tRNA aminoacylation by arginyl-tRNA synthetase: induced conformations during substrates binding," *The EMBO Journal*, vol. 19, no. 21, pp. 5599–5610, 2000.

[183] T. Kobayashi, T. Takimura, R. Sekine, V. P. Kelly, K. Vincent, K. Kamata, K. Sakamoto, S. Nishimura, and S. Yokoyama, "Structural snapshots of the KMSKS loop rearrangement for amino acid activation by bacterial tyrosyl-tRNA synthetase," *J. Mol. Biol.*, vol. 346, pp. 105–117, Feb 2005.

[184] K. Kumar, S. M. Woo, T. Siu, W. A. Cortopassi, F. Duarte, and R. S. Paton, "Cation-pi interactions in protein-ligand binding: theory and data-mining reveal different roles for lysine and arginine," *Chem. Sci.*, vol. 9, pp. 2655–2665, 2018.

[185] R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman, "Side-chain flexibility in proteins upon ligand binding," *Proteins*, vol. 39, pp. 261–268, May 2000.

[186] K. Gunasekaran and R. Nussinov, "How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding," *J. Mol. Biol.*, vol. 365, pp. 257–273, Jan 2007.

[187] S. Cusack, A. Yaremchuk, and M. Tukalo, "The crystal structure of the ternary complex of T. thermophilus seryl-tRNA synthetase with tRNA(Ser) and a seryl-adenylate analogue reveals a conformational switch in the active site," *EMBO J.*, vol. 15, pp. 2834–2842, Jun 1996.

[188] A. M. Gallina, P. Bork, and D. Bordo, "Structural analysis of protein-ligand interactions: the binding of endogenous compounds and of synthetic drugs," *J. Mol. Recognit.*, vol. 27, pp. 65–72, Feb 2014.

[189] N. P. Chowdhury, A. M. Mowafy, J. K. Demmer, V. Upadhyay, S. Koelzer, E. Jayamani, J. Kahnt, M. Hornung, U. Demmer, U. Ermler, and W. Buckel, "Studies on the mechanism of electron bifurcation catalyzed by electron transferring flavoprotein (Etf) and butyryl-CoA dehydrogenase (Bcd) of Acidaminococcus fermentans," *J. Biol. Chem.*, vol. 289, pp. 5145–5157, Feb 2014.

[190] S. Barelier, T. Sterling, M. J. O'Meara, and B. K. Shoichet, "The Recognition of Identical Ligands by Unrelated Proteins," *ACS Chem. Biol.*, vol. 10, pp. 2772–2784, Dec 2015.

[191] E. Schmitt, L. Moulinier, S. Fujiwara, T. Imanaka, J. C. Thierry, and D. Moras, "Crystal structure of aspartyl-tRNA synthetase from Pyrococcus kodakaraensis KOD: archaeon specificity and catalytic mechanism of adenylate formation," *EMBO J.*, vol. 17, pp. 5227–5237, Sep 1998.

[192] W. W. Navarre, S. B. Zou, H. Roy, J. L. Xie, A. Savchenko, A. Singer, E. Edvokimova, L. R. Prost, R. Kumar, M. Ibba, and F. C. Fang, "PoxA, yjeK, and elongation factor P coordinately modulate virulence and drug resistance in Salmonella enterica," *Mol. Cell*, vol. 39, pp. 209–221, Jul 2010.

[193] A. Vidal-Cros and H. Bedouelle, "Role of residue Glu152 in the discrimination between transfer RNAs by tyrosyl-tRNA synthetase from Bacillus stearothermophilus," *J. Mol. Biol.*, vol. 223, pp. 801–810, Feb 1992.

[194] Y. Xin, W. Li, D. S. Dwyer, and E. A. First, "Correlating amino acid conservation with function in tyrosyl-tRNA synthetase," *J. Mol. Biol.*, vol. 303, pp. 287–298, Oct 2000.

[195] F. Kaiser, A. Eisold, and D. Labudde, "A Novel Algorithm for Enhanced Structural Motif Matching in Proteins," *J. Comput. Biol.*, vol. 22, pp. 698–713, Jul 2015.

[196] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, Feb 2007.

[197] J. Desaphy, E. Raimbaud, P. Ducrot, and D. Rognan, "Encoding protein-ligand interaction patterns in fingerprints and graphs," *J Chem Inf Model*, vol. 53, pp. 623–637, Mar 2013.

[198] J. Konc and D. Janezic, "An improved branch and bound algorithm for the maximum clique problem," *MATCH Commun. Math. Comput. Chem.*, vol. 58, 2007.

[199] E. A. Coutsias, C. Seok, and K. A. Dill, "Using quaternions to calculate RMSD," *J Comput Chem*, vol. 25, pp. 1849–1857, Nov 2004.

[200] P. Liu, D. K. Agrafiotis, and D. L. Theobald, "Fast determination of the optimal rotational matrix for macromolecular superpositions," *J Comput Chem*, vol. 31, pp. 1561–1563, May 2010.

[201] A. R. Bradley, A. S. Rose, A. Pavelka, Y. Valasatava, J. M. Duarte, A. Prlic, and P. W. Rose, "MMTF – An efficient file format for the transmission, visualization, and analysis of macromolecular structures," *PLoS Comput. Biol.*, vol. 13, p. e1005575, Jun 2017.

[202] E. Akiva, S. Brown, D. E. Almonacid, A. E. Barber, A. F. Custer, M. A. Hicks, C. C. Huang, F. Lauck, S. T. Mashiyama, E. C. Meng, D. Mischel, J. H. Morris, S. Ojha, A. M. Schnoes, D. Stryke, J. M. Yunes, T. E. Ferrin, G. L. Holliday, and P. C. Babbitt, "The Structure-Function Linkage Database," *Nucleic Acids Res.*, vol. 42, pp. D521–530, Jan 2014.

[203] Y. Sato, I. Sagami, and T. Shimizu, "Critical role of the neuronal nitric-oxide synthase heme proximal side residue, Arg418, in catalysis and electron transfer," *J. Inorg. Biochem.*, vol. 87, pp. 261–266, Dec 2001.

[204] M. Biasini, "PV - webgl-based protein viewer," Nov 2014.

[205] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, (San Francisco, CA, USA), pp. 487–499, Morgan Kaufmann Publishers Inc., 1994.

[206] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data mining and knowledge discovery*, vol. 8, no. 1, pp. 53–87, 2004.

[207] B. Gärtner, "Fast and robust smallest enclosing balls," *Algorithms-ESA'99*, pp. 693–693, 1999.

[208] R. R. Sokal, "A statistical method for evaluating systematic relationships," *Univ Kans Sci Bull*, vol. 38, pp. 1409–1438, 1958.

[209] T. J. Wheeler and J. D. Kececioglu, "Multiple alignment by aligning alignments," *Bioinformatics*, vol. 23, pp. i559–568, Jul 2007.

[210] P. M. Morse, "Diatomic molecules according to the wave mechanics. II. vibrational levels," *Physical Review*, vol. 34, no. 1, p. 57, 1929.

[211] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[212] S. Schmitt, D. Kuhn, and G. Klebe, "A new method to detect related function among proteins independent of sequence and fold homology," *J. Mol. Biol.*, vol. 323, pp. 387–406, Oct 2002.

[213] G. Rhodes, *Crystallography made crystal clear: a guide for users of macromolecular models*. Elsevier, 2010.

[214] G. Yang, "The complexity of mining maximal frequent itemsets and maximal frequent patterns," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 344–353, ACM, 2004.

[215] P. J. Pereira, A. Bergner, S. Macedo-Ribeiro, R. Huber, G. Matschiner, H. Fritz, C. P. Sommerhoff, and W. Bode, "Human beta-tryptase is a ring-like tetramer with active sites facing a central pore," *Nature*, vol. 392, pp. 306–311, Mar 1998.

[216] Y. Xue, M. Okvist, O. Hansson, and S. Young, "Crystal structure of spinach plastocyanin at 1.7 Å resolution," *Protein Sci.*, vol. 7, pp. 2099–2105, Oct 1998.

[217] S. Young, K. Sigfridsson, K. Olesen, and O. Hansson, "The involvement of the two acidic patches of spinach plastocyanin in the reaction with photosystem I," *Biochim. Biophys. Acta*, vol. 1322, pp. 106–114, Dec 1997.

[218] H. J. Feldman and P. Labute, "Pocket similarity: are alpha carbons enough?," *J Chem Inf Model*, vol. 50, pp. 1466–1475, Aug 2010.

[219] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 37, pp. 205–211, Apr 1951.

[220] B. Y. Chen, D. H. Bryant, A. E. Cruess, J. H. Bylund, V. Y. Fofanov, D. M. Kristensen, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "Composite motifs integrating multiple protein structures increase sensitivity for function prediction," *Comput Syst Bioinformatics Conf*, vol. 6, pp. 343–355, 2007.

[221] Y. Valasatava, A. R. Bradley, A. S. Rose, J. M. Duarte, A. Prlic, and P. W. Rose, "Towards an efficient compression of 3D coordinates of macromolecular structures," *PLoS ONE*, vol. 12, no. 3, p. e0174846, 2017.

[222] V. J. Haupt, S. Daminelli, and M. Schroeder, "Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key," *PLoS ONE*, vol. 8, no. 6, p. e65894, 2013.

[223] M. Brylinski and W. P. Feinstein, "eFindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands," *J. Comput. Aided Mol. Des.*, vol. 27, pp. 551–567, Jun 2013.

[224] D. A. Kirshner, J. P. Nilmeier, and F. C. Lightstone, "Catalytic site identification–a web server to identify catalytic site structural matches throughout PDB," *Nucleic Acids Res.*, vol. 41, pp. W256–265, Jul 2013.

[225] L. Xie, L. Xie, and P. E. Bourne, "A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery," *Bioinformatics*, vol. 25, pp. i305–312, Jun 2009.

[226] M. Hendlich, F. Rippmann, and G. Barnickel, "LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins," *J. Mol. Graph. Model.*, vol. 15, pp. 359–363, Dec 1997.

[227] A. C. Wallace, N. Borkakoti, and J. M. Thornton, "TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites," *Protein Sci.*, vol. 6, pp. 2308–2323, Nov 1997.

[228] A. Stark and R. B. Russell, "Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures," *Nucleic Acids Res.*, vol. 31, pp. 3341–3344, Jul 2003.

[229] R. L. Baldwin, "Making a network of hydrophobic clusters," *Science*, vol. 295, pp. 1657–1658, Mar 2002.

[230] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annu Rev Biophys*, vol. 37, pp. 289–316, 2008.

[231] S. Selvaraj and M. M. Gromiha, "Importance of hydrophobic cluster formation through long-range contacts in the folding transition state of two-state proteins," *Proteins*, vol. 55, pp. 1023–1035, Jun 2004.

[232] V. Frappier, M. Chartier, and R. J. Najmanovich, "ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability," *Nucleic Acids Res.*, vol. 43, pp. 395–400, Jul 2015.

[233] A. K. Hirsch, F. R. Fischer, and F. Diederich, "Phosphate recognition in structural biology," *Angew. Chem. Int. Ed. Engl.*, vol. 46, no. 3, pp. 338–352, 2007.

[234] K. A. Denessiouk, M. S. Johnson, and A. I. Denesyuk, "Novel CalphaNN structural motif for protein recognition of phosphate ions," *J. Mol. Biol.*, vol. 345, pp. 611–629, Jan 2005.

[235] B. W. Matthews, "Protein-DNA interaction. No code for recognition," *Nature*, vol. 335, pp. 294–295, Sep 1988.

[236] J. Skolnick, M. Gao, A. Roy, B. Srinivasan, and H. Zhou, "Implications of the small number of distinct ligand binding pockets in proteins for drug discovery, evolution and biochemical function," *Bioorg. Med. Chem. Lett.*, vol. 25, pp. 1163–1170, Mar 2015.

[237] Z. Zhao, L. Xie, L. Xie, and P. E. Bourne, "Delineation of Polypharmacology across the Human Structural Kinome Using a Functional Site Interaction Fingerprint Approach," *J. Med. Chem.*, vol. 59, pp. 4326–4341, May 2016.

[238] D. E. Koshland Jr, "The key-lock theory and the induced fit theory," *Angewandte Chemie International Edition in English*, vol. 33, no. 23-24, pp. 2375–2378, 1995.

[239] H. Qin, L. Lim, and J. Song, "Protein dynamics at Eph receptor-ligand interfaces as revealed by crystallography, NMR and MD simulations," *BMC Biophys*, vol. 5, p. 2, Jan 2012.

[240] B. Ma, S. Kumar, C. J. Tsai, and R. Nussinov, "Folding funnels and binding mechanisms," *Protein Eng.*, vol. 12, pp. 713–720, Sep 1999.

[241] R. Nussinov and B. Ma, "Protein dynamics and conformational selection in bidirectional signal transduction," *BMC Biol.*, vol. 10, p. 2, Jan 2012.

[242] D. Joseph, G. A. Petsko, and M. Karplus, "Anatomy of a conformational change: hinged 'lid' motion of the triosephosphate isomerase loop," *Science*, vol. 249, pp. 1425–1428, Sep 1990.

[243] H. A. Carlson, "Protein flexibility is an important component of structure-based drug discovery," *Curr. Pharm. Des.*, vol. 8, no. 17, pp. 1571–1578, 2002.

[244] A. Gutteridge and J. Thornton, "Conformational changes observed in enzyme crystal structures upon substrate binding," *J. Mol. Biol.*, vol. 346, pp. 21–28, Feb 2005.

[245] J. F. Gibrat, T. Madej, and S. H. Bryant, "Surprising similarities in structure comparison," *Curr. Opin. Struct. Biol.*, vol. 6, pp. 377–385, Jun 1996.

[246] J. P. A. Moraes, G. L. Pappa, D. E. V. Pires, and S. C. Izidoro, "GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms," *Nucleic Acids Res.*, Apr 2017.

[247] J. M. Duarte, A. Srebniak, M. A. Scharer, and G. Capitani, "Protein interface classification by evolutionary analysis," *BMC Bioinformatics*, vol. 13, p. 334, Dec 2012.

[248] R. S. Kaminsky, N. Snavely, S. M. Seitz, and R. Szeliski, "Alignment of 3D point clouds to overhead images," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 63–70, Jun 2009.

[249] M. Pauly, N. J. Mitra, J. Wallner, H. Pottmann, and L. J. Guibas, "Discovering structural regularity in 3D geometry," *ACM Trans. Graph.*, vol. 27, pp. 43:1–43:11, Aug 2008.

GLOSSARY

alpha carbon

The primary carbon atom of the backbone of an amino acid.

angle θ

The angle between the side chains of two residues, defined by abstracting each side chain as a vector between alpha carbon and the most distant carbon side chain atom.

Arginine Tweezers

A unique structural motif in Class II aminoacyl-tRNA synthetases. The motif consists of two arginine residues, grasping the ligand similar to a pair of tweezers.

Backbone Brackets

A unique structural motif in Class I aminoacyl-tRNA synthetases. The motif consists of two residues that are not conserved in sequence. These residues mediate interactions with the ATP ligand via backbone hydrogen bonds.

beta carbon

The first carbon atom of the side chain of an amino acid. Glycine lacks the beta carbon atom.

computational structural motif detection

Computational structural motif detection is used to discover recurrent, evolutionarily conserved, and/or functionally important spatial residue patterns in macromolecular structures (e.g. proteins, DNA, or RNA). The problem of computational structural motif detection can be further categorized into template-based or template-free.

EC:x.x.x.x

The nomenclature used in this thesis to describe a four-level Enzyme Commission [41] identifier.

Fit3D

Fit3D is a collection of algorithms and implementations thereof for the template-based and template-free detection of structural motifs. In contrast to most competitors, Fit3D supports the computational representation of structural motifs at the atomic level and isofunctional mutations.

ligand

In the context of the thesis a small molecule (e.g. a drug compound) or ion that is specifically bound by a macromolecular structure such as a protein.

M1

The binding mode in structures of aminoacyl-tRNA synthetases where an ATP ligand is bound.

M2

The binding mode in structures of aminoacyl-tRNA synthetases where no ATP ligand is bound.

PDB:xxxx

The nomenclature used in this thesis to describe a four-character Protein Data Bank [38] identifier.

Pfam:PFxxxxx

The nomenclature used in this thesis to describe a Pfam [132] identifier.

REST

An application programming interface provided by a web service. The Protein Data Bank [38] allows the computational query of their search services via a REST endpoint.

RNA world hypothesis

A popular hypothesis that assumes early life was entirely based on RNA.

Rodin-Ohno hypothesis

A hypothesis formulated by RODIN AND OHNO in 1995 [18]. It states that ancient forms of Class I and Class II aminoacyl-tRNA synthetases were once coded complementary on the strands of a single gene ("Urzyme") in a bidirectional way.

SFLD:x

The nomenclature used in this thesis to describe a SFLD [202] identifier.

TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Division of Bioinformatics, Biotechnology Center TU Dresden

STATEMENT OF AUTHORSHIP

TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Division of Bioinformatics, Biotechnology Center TU Dresden

I hereby certify that I have authored this Dissertation entitled *Structural Bioinformatics to Understand the Origin of the Genetic Code* independently and without undue assistance from third parties. No other than the resources and references indicated in this thesis have been used. I have marked both literal and accordingly adopted quotations as such. There were no additional persons involved in the intellectual preparation of the present thesis. I am aware that violations of this declaration may lead to subsequent withdrawal of the degree.

Dresden, 26th June 2018

Florian Kaiser